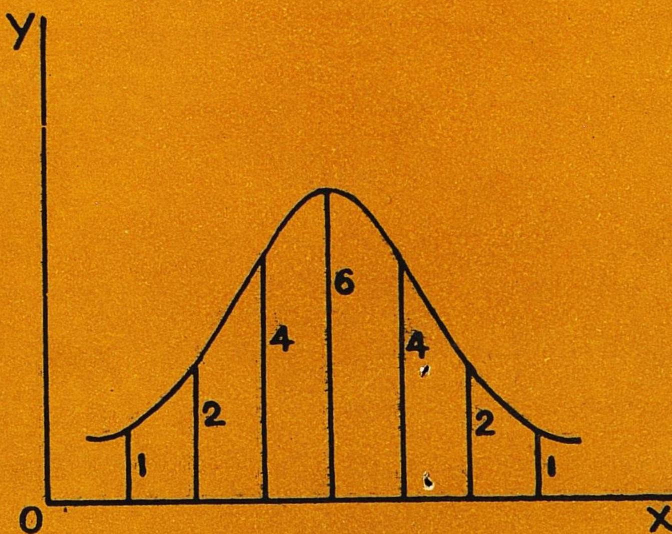
STATISTICS

HIGHER SECONDARY
FIRST YEAR

VOL. II



TAMILNADU TEXTBOOK SOCIETY

STATISTICS

Vol. II

Higher Secondary – First Year

TAMILNADU TEXTBOOK SOCIETY
MADRAS

First Edition—1980

Editorial Board Chairman
(Author & Review Committee Members)

Thiru. **M. Sankaranarayanan,** M.A., B.Sc.,
Joint Director of Statistics,
Department of Statistics,
MADRAS - 600 006.

REVIEW COMMITTEE MEMBERS:

Thiru. **T. K. Manickavachagam Pillai,** M.A., L.T.,
Professor of Mathematics (Retd.)
A.C. College of Technology;
MADRAS - 600 035.

Thiru. **R. Hanumantha Rao,** M.A.,
Professor of Mathematics,
P.S.G. Arts College,
COIMBATORE.

Price: Rs. 8-00

This book has been printed on concessional paper of 60 G.S.M. substance made available by the Government of India.

Printed at
SANKAR PRINTERS, MADRAS-600 018.

CONTENTS

FIRST YEAR :: SECOND PAPER

CHAPTER I

MEASURES OF CENTRAL TENDENCIES

We have so far considered how a large number of statistical data can be condensed by means of tables and represented in charts and graphs for easy understanding and comparing. But tables, charts and graphs have their own limitations. Various distributions in the form of tables cannot be compared directly.

Suppose we have two tables showing the wages' distribution of workers in two different factories. It may not be possible to have a definite conclusion by means of direct comparison of the data. In order to make the comparison easy and effective so as to arrive at a conclusion, we should have a common measure which should describe the characteristics of the given data. It may happen in any distribution that a few values may occur more frequently and a few may occur less frequently. The values which may occur more frequently may lie in a particular part or position of the distribution. In most cases the particular part or position may be the central part or central position and hence that value may be taken as the central value for that distribution. The data in this distribution may have a tendency either to be equal to the central value or to tend towards that central value. Hence, that central value may be taken as a measure of central tendency. As this measure indicates location in the distribution this measure is also known as a **Measure of Location.**

Let us select 100 uniform plots with measurements 10 $m \times 10m$ in different parts of a taluk and harvest the paddy crop and record the yield obtained from each of the plots. We are sure that the yields of all the 100 plots may not be equal to one another and the yield varies from plot to plot. There may be some extreme high yield due to better application of fertilisers and other inputs.

Similarly, in a few cases the yeild may be very poor due to want of proper irrigation facilities or due to pest attack. However, we may find that the yields in the remaining cases, other than the extreme values, may not differ much from one another or they may be close to one another. In other words, we can say that they cluster around some particular value or they are tending towards a particular value. The values will have a tendency to cluster around a particular value. Hence, the particular value around which the other values cluster may be called a **central value**. Such a central value is known as a Measure of Central tendency or measure of location. For different frequency distributions there may be different values of central tendency. Therefore, it can be said that frequency distributions may differ from one another in this aspect, namely in the measure of central tendency.

There are different measures of central tendency. They are Mean or Average, Median and Mode. The mean may be classified into Arithmetic Mean, Geometric Mean and Harmonic Mean. The Arithmetic Mean may be either of two categories namely Simple Mean and Weighted Mean.

Characteristics of a good statistical average

Though there are different types of averages, each average has its own advantages and disadvantages and hence different averages are used on different occasions. No single average is suitable for all purposes. We have to select the best for the occasion. The average which satisfies certain characteristics can be considered as the best.

1. It should be capable of being calculated by a well defined mathematical formula.
2. It should be based on the values of all the items in the distribution.
3. The value of the average should not be unduly affected by extreme high or extreme low values. In other words, the value of the average should not be altered by a wide range because of the extreme values.
4. The computation of the average should be simple and easy for understanding.

5. It should be amenable for any algebraic treatment.
6. It should be stable. The value of the averages should not be affected much by small changes in the method of computation.

Arithmetic Mean (or) Mean (or) Average

Arithmetic Mean or Mean or Average are one and the same. Generally, the term Mean is always used in Statistics. The term 'Average' is a familiar one and its computation is also much familiar and easy. Mean can be calculated for (1) ungrouped data; (2) discrete frequency distributions; and (3) continous frequency distributions.

A distribution may consist of a number of units or individuals. For calculating the average, the values of the individuals are added and their total value is first computed. The total value thus arrived at, will be equally divided by the number of units or individuals and the average value is arrived at. This is called the simple average or arithmetic mean or simple mean. This method is called 'direct method.'

Let us consider the following example:

The weight of five bundles are given in *kg*. Find the average weight of one bundle.

Let x represent the weight of the bundle and the numbers 1, 2, 3, 4 and 5 represent the serial numbers of the bundles.

Weight in symbol	Actual weight kg.
x_1	45
x_2	50
x_3	65
x_4	75
x_5	40

$$\text{Average} = \frac{45 + 50 + 65 + 75 + 40}{5}$$

$$= \frac{275}{5} = 55 \text{ kg.}$$

If x represents the value of the variable, the average value is generally written as $\bar{x}$. The total number of units or individuals may be represented by 'n' or 'N'.

If x_1, x_2, x_3, ...x_n are the values of 'n' different units, then the average $\bar{x}$ will be written as follows. In this problem 'n' is equal to 5.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \ldots\ldots x_n}{n}$$

(whether it is 'N' or 'n' is immaterial at this stage).

The letter 'S' stands for the term Summation which means addition. In Greek, the letter for the term summation is written as Σ (called sigma). The first term will be written at the bottom and the last term will be written at the top of this letter $\sum\limits_{1}^{n}$. Therefore the average can be written as follows:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \quad \text{or} \quad 1/n \sum_{i=1}^{n} x_i$$

where 'i' ranging from 1 to n represents the imaginery unit and there are 'n' such imaginery units in the distribution. Since 'n' is a constant number it has been taken outside the Σ symbol.

When the total value of all the items is divided by the number of items or units or individuals, the average will be obtained.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\text{Mean} = \frac{\text{Total value of all the items } \left(\sum_{i=1}^{n} x_i\right)}{\text{The total number of items } (n)}$$

In other words, when the average is multiplied by the number of items or units or individuals, we will get the total value of all the items.

$$n \times \bar{x} = \sum_{i=1}^{n} x_i \quad \text{or} \quad n\bar{x} = \sum_{i=1}^{n} x_i \quad \text{or} \quad \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

Property of the Arithmetic Mean

Some of the values of the units or members of a distribution may be greater than the average and some others may be less than the average. Therefore, the difference between the value and the average will be either a positive quantity (+) or a negative quantity (—)depending upon the value of each unit when compared with the mean. These differences are generally called deviations or variations. The sum of the deviations of all the items from the Mean will always be '0'. This is an important property of the Arithmetic Mean. This can be proved as follows:

Example

Value x_i (1)	Deviation $(di) = (x_i - \bar{x})$ (2)
45	45 — 55 = —10
50	50 — 55 = — 5
65	65 — 55 = +10
75	75 — 55 = +20
40	40 — 55 = —15
275	0

$$\text{Average} = \frac{275}{5} = 55.$$

Let us take column 2 separately and examine it by splitting into two portions or sets as follows:

	First set		Second set
	45	(—)	55
	50	(—)	55
	65	(—)	55
	75	(—)	55
	40	(—)	55
Total	275	(—)	275 = 0.

Add these two sets of values.

The total of the first set of figures will be equal to the total of all the items. The total of the second set of figures will be equal to the product of the average multiplied by the number of units which will be equal to the total value of all the items. Hence the difference is '0'.

We can now consider the formula.

Value (x_i) (1)	Deviation (d_i) (2)
x_1	$x_1 - \bar{x} = d_1$
x_2	$x_2 - \bar{x} = d_2$
x_3	$x_3 - \bar{x} = d_3$
x_4	$x_4 - \bar{x} = d_4$
x_5	$x_5 - \bar{x} = d_5$
x_n	$x_n - \bar{x} = d_n$

$$d_1 + d_2 + d_3 + d_4 + d_5 \ldots. + d_n = \sum_{i=1}^{n} d_i$$

$$= (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + (x_4 - \bar{x}) + (x_5 - \bar{x}) \ldots\ldots + (x_n - \bar{x})$$

Remove the brackets and then re-arrange them by grouping the positive items and negative items.

$$x_1 - \bar{x} + x_2 - \bar{x} + x_3 - \bar{x} + x_4 - \bar{x} + x_5 - \bar{x} \ldots. + x_n - \bar{x}.$$

$$= (x_1 + x_2 + x_3 + x_4 + \ldots + x_n) - (\bar{x} + \bar{x} + \bar{x} + \bar{x} + \bar{x} + \ldots.\ n \text{ times})$$

$$= \Sigma\,(x_1 - n\bar{x}) = 0$$

$$\therefore \text{General formula} = \sum_{i=1}^{n} x_i - n\bar{x} = 0.$$

This is a very important property of the Arithmetic Mean.

Computation of Arithmetic Mean

There are different methods for the computation of the Arithmetic Mean. All these methods are developed to save time and also to avoid unnecessary labour.

The following is the monthly wages of 10 workers in a factory. Let us calculate the average wage of the workers.

Method I

S.No. of the workers	Wages (Rs.)
1	115
2	112
3	117
4	118
5	111
6	115
7	112
8	119
9	111
10	110
Total	1140

$$\text{Average wage per worker} = \frac{1140}{10} = \text{Rs. 114 per head.}$$

Let us examine the wages of the above workers once again. We notice that the wages of a few workers are the same. In other words, the same value is repeated or frequented more than once. We should now re-write the above table.

Method II

Wage per worker (Rs.) (1)	No. of workers (2)	Total wages (Rs.) (col.1 × col. 2) (3)
115	2	230
112	2	224
117	1	117
118	1	118
111	2	222
119	1	119
110	1	110
Total	10	1140

$$\text{Total wages} = \text{Rs. } 1140.$$

$$\text{Average wage} = \frac{1140}{10} = \text{Rs. } 114 \text{ per head.}$$

In the above two methods, the average wage per worker is the same. In the first method, which is a direct method, we have added straight away all the values and then calculated the average. In the second method, we have calculated the frequencies of each of the values as given in column 2. Afterwards we have multiplied the wages by the frequencies and given the total wages of the workers under the particular group as given in column (3). The total wages of all the workers as given in column 3 is divided by the total number of workers and the average wage is calculated. It shall be noted that calculation of frequencies for each value amounts

to classification and preparation of a frequency table. But the frequency table prepared on this basis is without the various class intervals. Instead of preparing each class with their class limits, we have the various classes with more or less their mid-values. The above method is also a short and direct method.

Method III

In the above example, we find that the wages of all the workers is more than Rs. 100. Hence we can subtract Rs. 100/- from the wages of each worker and the balance can be written as follows:

S.No. of the workers (1)	Actual wages - Rs. 100 (2)
1	15
2	12
3	17
4	18
5	11
6	15
7	12
8	19
9	11
10	10
Total	140

$$\text{Average} = \frac{140}{10} = \text{Rs. } 14 \text{ per head.}$$

Since we have subtracted Rs. 100 from the wages of each of the workers and calculated the average, we should add Rs. 100/- to this new assumed average wage of Rs. 14/- and find out the average wage (in original value) of the workers.

Average wage = 14 + 100 = Rs. 114 per head.
(in original value)

We can compare this with the situation where workers are getting their salary in terms of one rupee currencies. Since the number of one rupee notes are very large, it takes time for counting then and it is also difficult to carry in hand, the workers would like to have one 100 rupee note and the balance in terms of one rupee note. But in the house they require only smaller denominations of the currency namely one rupee note for their day to day expense. Therefore, they may again exchange the 100 rupees note into one rupee notes.

In order to have a clear distinction between the two averages, we can consider the original values as X and the new values obtained after subtraction of 100 from each as Y. Then the formula will be as follows:

$$\text{Let } Y = X - 100;$$

$$\therefore \bar{Y} = \bar{X} - 100$$

$$\therefore \bar{Y} + 100 = \bar{X}.$$

The above simplification can be followed in the case of method II.

Example:

We can consider 118 as an arbitrary value. The value of 'd' will be equal to $x - 118$.

$d = x - 118$	No. of workers	Total value
$115 - 118 = -3$	2	-6
$112 - 118 = -6$	2	-12
$117 - 118 = -1$	1	-1
$118 - 118 = 0$	1	0
$111 - 118 = -7$	2	-14
$119 - 118 = +1$	1	1
$110 - 118 = -8$	1	-8
Total	10	-40

$$\text{Average of } `d' = \bar{d} = \frac{-40}{10} = -4$$

ie: $d = x - 118, \quad \therefore \ \bar{d} = \bar{x} - 118$

$$\begin{aligned}\bar{x} &= \bar{d} + 118 \\ &= -4 + 118 = 114 \\ &= \text{Rs. } 114.\end{aligned}$$

Instead of taking a value occupying the central position (118) we can take an other value 115. The position would emerge as follows:

$d = x - 115$	No. of workers	Total value
$115 - 115 = 0$	2	0
$112 - 115 = -3$	2	-6
$117 - 115 = 2$	1	2
$118 - 115 = 3$	1	3
$111 - 115 = -4$	2	-8
$119 - 115 = 4$	1	4
$110 - 115 = -5$	1	-5
Total	10	-10

$$\bar{d} = \frac{-10}{10} = -1$$

$$\therefore \quad \bar{x} = \bar{d} + 115$$

$$= -1 + 115 = 114.$$

Though we have adopted different arbitrary values, the method adopted in all the cases are one and the same and the final result of the average is also the same.

There are many shorter methods and we shall examine the advantages of these methods. Let us examine the salary of the following 10 workers and compute their average.

Short-cut Method I

S. No. of the workers (1)	Salary (x) Rs. (2)	Short cut Method (A=155) 'd' = x– 155 (3)
1	135	135 — 155 = —20
2	145	145 — 155 = – 10
3	180	180 — 155 = 25
4	185	185 — 155 = 30
5	195	195 — 155 = 40
6	155	155 — 155 = 0
7	170	170 — 155 = 15
8	130	130 — 155 = —25
9	140	140 — 155 = – 15
10	165	165 — 155 = 10
Total	1600	50

$$\text{Average} = \bar{d} = \frac{\Sigma d}{n}$$

$$= \frac{50}{10} = 5.$$

$$\therefore \quad \bar{x} = \bar{d} + A$$

$$= 5+155 = \text{Rs. } 160.$$

Since we have reduced each of the original values by subtracting 155 from each, we have to add 155 to the average of the new values to get the average of the original values. The operation is just reverse from the original operation.

The total salary of all the 10 persons as per column 2, ie. by direct method is Rs. 1600. Hence the average salary as per direct method is also Rs. 160. In both the methods, the average obtained is the same.

This method of computation is known as computation of Arithmetic Mean by shifting the base since we are shifting the base to 155.

General case

Values xi	Deviation from A di
x_1	$x_1 - A$
x_2	$x_2 - A$
x_3	$x_3 - A$
..	..
x_n	$x_n - A$

Sum of the deviation

$$\sum_{i=1}^{n} d_i = (x_1 - A) + (x_2 - A) + \dots (x_n - A)$$
$$= x_1 + x_2 + \dots + x_n - nA$$
$$= \Sigma x - nA$$
$$= n\bar{x} - nA$$

$$\frac{\Sigma d_i}{n} = \frac{n\bar{x}}{n} - \frac{nA}{n}$$
$$\bar{d} = \bar{x} - A$$
$$\bar{x} = \bar{d} + A$$

Short-cut Method II

In this method we divide each value by a common number instead of subtracting an arbitrary value (155) from each value as adopted in the first method. Let us consider the same example and divide the salary of each person by a common number 5.

S.No. of the person (1)	Salary (X) Rs. (2)	$d = x/5$ (3)
1	135	$135 \div 5 = 27$
2	145	$145 \div 5 = 29$
3	180	$180 \div 5 = 36$
4	185	$185 \div 5 = 37$
5	195	$195 \div 5 = 39$
6	155	$155 \div 5 = 31$
7	170	$170 \div 5 = 34$
8	130	$130 \div 5 = 26$
9	140	$140 \div 5 = 28$
10	165	$165 \div 5 = 33$
Total	1600	Total = 320

Total of 'd' $= \Sigma d = 320.$

$$\text{Average of } d = \bar{d} = \frac{\Sigma d}{n} = \frac{320}{10} = 32.$$

$$\text{We have assumed } d = \frac{x}{5}$$

$\therefore \quad 5d = x; \qquad$ ie; $x = 5d.$

$\therefore \quad \bar{x} = 5 \times \bar{d} = 5 \times 32 =$ Rs. 160.

This method is known as computation by change of scale. This method can be well compared with our day-to-day experience. We can consider the workers getting their salary in terms of one rupee notes. The number of one rupee notes can be exchanged for 5 rupees notes. The number of one rupee notes each worker would get is the same as the value given in Col (2). These one rupee notes can be exchanged for five rupees notes and the number of five rupees notes each worker would get is given in Column (3). The average number of five rupees notes each one would get is 32 since the total number of five rupees notes is 320. These five rupees notes can be exchanged for one rupee notes again. As each five rupees note can be converted into 5 one rupee notes, the 32 five rupees notes can be changed into (32×5) 160 one rupee notes.

Since we have initially reduced the size of the value by dividing each value by 5, we have to multiply the average of the new value by 5. The process adopted is just reversal of the original operation.

Short-cut Method III Computation by shifting the base and changing the scale

There is still another short cut method which involves the previous two methods simultaneously. In this method we subtract first an arbitrary value from each one and the balance is divided by a common value. However this will become cumbersome if there is no common multiple.

S. No. (1)	Salary (2)	$y = x - 100$ (3)	$z = \frac{(x-100)}{5} = \frac{y}{5}$ (4)
1	135	35	7
2	145	45	9
3	180	80	16
4	185	85	17
5	195	95	19
6	155	55	11
7	170	70	14
8	130	30	6
9	140	40	8
10	165	65	13
Total	1600	600	120

$$\text{Average} = \frac{1600}{10} \qquad \frac{600}{10} \qquad \frac{120}{10}$$

$$\text{i.e.} \quad \bar{x} = 160 \qquad \bar{y} = 60 \qquad \bar{z} = 12$$

$$\bar{x} = \bar{y} + 100$$

$$\bar{x} = 60 + 100 \qquad \bar{x} = 5\bar{z} + 100$$

$$= 160 \qquad = 5 \times 12 + 100$$

$$= 60 + 100$$

$$= 160$$

As in the case of Method I, we have subtracted 100 from each of the 'x' values and the values thus arrived at are given in column (3) as 'y'. values. Then each of the 'y' values given in column (3) is divided by 5 to reduce it still further as in the case of Method II and the values thus obtained are given in the column (4) as 'z' values.

In the process of conversion from 'x' values to 'z' values the following two operations are done.

1. Subtraction of 100 from 'x' to convert into 'y'.
2. Division of 'y' by 5 to get 'z'.

$$z = \frac{y}{5} = \frac{x - 100}{5} \; ; \quad \bar{z} = \frac{\bar{y}}{5} = \frac{\bar{x} - 100}{5}$$

$$\therefore \; 5\,\bar{z} = \bar{x} - 100.$$

$$\bar{x} = 5\bar{z} + 100$$

$$= 5 \times 12 + 100$$

$$= 60 + 100 = 160 = \textbf{Rs. 160.}$$

Since division by 5 is the last step, we have to multiply the average of Z by 5. Since the subtraction of 100 from 'x' is the first step, we have to add 100 to the average of 'y' obtained by multiplying the average of Z by 5. The operations are just opposite to the ones carried out in the beginning but they are now carried out in the reverse order.]

Generally the third method, which is the combination of the first two methods will be adopted. Further, the arbitrary value will be denoted by the letter 'A' and the diviser is denoted by the letter 'C'. The formula would emerge as follows:

$$d = \frac{x - A}{C}$$

$$\bar{d} = \frac{\bar{x} - A}{C}$$

$$c.\bar{d}. = \bar{x} - A$$

$$\therefore \; \bar{x} = c.\,\bar{d} + A$$

Mean of a Discrete Frequency Distribution

(i) Direct Method

We shall consider the following frequency distribution;

Value (X) kg.	Frequency f	Total Value (xf) or (fx)
x_1 15	f_1 3	($x_1 f_1$) 45
x_2 12	f_2 4	($x_2 f_2$) 48
x_3 17	f_3 2	($x_3 f_3$) 34
x_4 18	f_4 1	($x_4 f_4$) 18
	Total 10	145

Since there are four values of x, they are denoted by the symbols x_1, x_2 x_3 and x_4 and their repective frequencies (f) are denoted by the symbols f_1, f_2, f_3 and f_4.

The total of each value (xf) is obtained by multiplying the value by its frequency. Thus there are four totals represented by $x_1 f_1$, $x_2 f_2$, $x_3 f_3$, and $x_4 f_4$. Their grand total is equal to $x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4$. The total number of items is equal to the total of the frequencies, that is, $f_1 + f_2 + f_3 + f_4$.

$$\text{The average } \bar{x} = \frac{\text{The grand total of the values}}{\text{Total number of items.}}$$

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4}{f_1 + f_2 + f_3 + f_4}$$

The formula can be expressed by the following symbol.

$$\bar{x} = \frac{\Sigma\, x.f}{\Sigma f} \text{ (or) } \frac{\Sigma fx}{\Sigma f}$$

$$\bar{x} = \frac{45 + 48 + 34 + 18}{3+4+2+1} = \frac{145}{10} = \mathbf{14.5 \text{ kg.}}$$

Short-cut Method — I (Shiftihg the base)

We can adopt short cut method also. We can assume 10 as the arbitrary value. Then the values will be changed as follows without any change in frequencies:

'd'	f	df.
15 — 10 = 5	3	15
12 — 10 = 2	4	8
17 — 10 = 7	2	14
18 — 10 = 8	1	8
	10	45

Average $\bar{d} = \frac{45}{10} = 4.5$

$d = x - 10.$

$\bar{d} = \bar{x} - 10$

$\therefore \quad \bar{x} = \bar{d} + 10 = 4.5 + 10 = 14.5.$ kg.

Shortcut Method II: (Changing the scale)

In this method, each of the discrete values can be reduced in size by dividing it by a constant number.

x kg. (1)	f (2)	xf (3)	$d = x/15$ (4)	$d.f.$ (5)
60	2	120	4	8
75	4	300	5	20
90	6	540	6	36
105	3	315	7	21
135	5	675	9	45
Total	20	1950		130

$$\Sigma xf = 1950; \quad \bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{1950}{20} = 97.5 \text{ kg.}$$

In the above case, each value can be divided by 15 and the value obtained is given in col(4). The product of 'd' and the frequency is given in col. (5).

$$\Sigma fd \text{ (or) } \Sigma df = 130. \quad \therefore \; \bar{d} = \frac{130}{20} = 6.5$$

$$d = x/15; \quad \therefore 15\,d = x; \therefore 15\,\bar{d} = \bar{x}\,; \quad \therefore \bar{x} = 15 \times 6.5 = 97.5$$

General Case

xi	fi	$d_i = \dfrac{xi}{c}$	fi di
x_1	f_1	$d_1 = \dfrac{x_1}{c}$	$f_1 \; \dfrac{x_1}{c} = f_1\, d_1$
x_2	f_2	$d_2 = \dfrac{x_2}{c}$	$f_2 \; \dfrac{x_2}{c} = f_2\, d_2$
x_n	f_n	$d_n = \dfrac{x_n}{c}$	$f_n \; \dfrac{x_n}{c} = f_n\, d_n$

$$\begin{aligned} \Sigma f_i\, d_i &= f_1 \frac{(x_1)}{c} + f_2 \frac{(x_2)}{c} + \ldots f_n \frac{(x_n)}{c} \\ &= 1/c\,(f_1 x_1 + f_2 x_2 \ldots\ldots + f_n x_n) \\ &= 1/c \;\Sigma f_i\, x_i \end{aligned}$$

Dividing by N

$$\frac{\Sigma f_i\, d_i}{N} = 1/c \;\frac{\Sigma f_i\, x_i}{N}$$

$$\bar{d} = \frac{\bar{x}}{c}$$

$$c\bar{d} = \bar{x}$$

We find that the averages obtained both by direct method and short-cut method are one and the same.

Short-cut Method III: (Shifting the base and changing the scale)

Let us combine the shortcut methods I and II for calculation of the averages. We shall consider the same example. We shall take 90 as the arbitrary value and subtract it from each value of x and the result obtained can be divided by a common number 15.

x (1)	f (2)	$y = x - 90$ (3)	$d = y/15$ (4)	fd (5)
60	2	$60 - 90 = -30$	-2	-4
75	4	$75 - 90 = -15$	-1	-4
90	6	$90 - 90 = 0$	0	0
105	3	$105 - 90 = 15$	1	3
135	5	$135 - 90 = 45$	3	15
	20			10

The value of 'd' obtained from the value of 'x' by the substitution $d = \dfrac{x - 90}{15}$ is given in col (4). The product of each item in column (4) and the respective frequency is given in column (5).

$$\Sigma fd = 10; \qquad \Sigma f = 20$$

$$\bar{d} = 10/20 = 0.5$$

$$\frac{x - 90}{15} = d\,; \quad \frac{\bar{x} - 90}{15} = \bar{d}$$

$$15\,\bar{d} = \bar{x} - 90$$

$$\therefore \bar{x} = 15\,\bar{d} + 90$$
$$= 15 \times 0.5 + 90$$
$$= 7.5 + 90$$
$$= 97.5$$

General Case:

x	f	$d = \dfrac{x - A}{c}$	fd
x_1	f_1	$d1 = \dfrac{x_1 - A}{c}$	$f1d1$
x_2	f_2	$d2 = \dfrac{x_2 - A}{c}$	$f2d2$
xn	fn	$dn = \dfrac{xn - A}{c}$	$fndn$

$$\Sigma fidi = f1 \frac{(x1 - A)}{c} + f2 \frac{(x2 - A)}{c} + \ldots + fn \frac{(xn - A)}{c}$$
$$= 1/c\,(f1x1 - f1A + f2x2 - f2A + \ldots\ldots fnxn - fnA)$$
$$= 1/c\,[(f1x1 + f2x2 \ldots + fnxn) - (f1A + f2A + \ldots fnA)\}]$$
$$= 1/c\,\{(\Sigma xifi - A\,\Sigma fi)\}$$
$$= 1/c\,(\Sigma xifi - NA)$$
$$= 1/c\,(N\bar{x} - NA)$$
$$= \frac{N}{C}(\bar{x} - A)$$

Dividing by N

$$\frac{\Sigma fidi}{N} = \frac{N}{C}\left\{\frac{(\bar{x} - A)}{N}\right\}$$
$$\bar{d} = \frac{\bar{x} - A}{C}$$
$$c\bar{d} = \bar{x} - A$$
$$c\bar{d} + A = \bar{x}$$

It is seen that in the case of frequency distribution of discrete values also, we can adopt shortcut methods with greater advantages. In all these shortcut methods, we reduce the size of the original values by a substitution and thus save time and labour in the computation.

Mean of a continuous frequency distribution

Direct Method

Let us consider the following continuous distribution, which is constructed from the raw data.

Class (1)	Frequency (2)
kg.	
60.5 — 70.5	1
70.5 — 80.5	5
80.5 — 90.5	9
90.5 — 100.5	14
100.5 — 110.5	15
110.5 — 120.5	4
120.5 — 130.5	2
Total	50

In the case of continuous frequency distribution giving the different classes and their resepective frequencies, we have to first calculate the class marks or the mid-values of each class. This is a very important step and this should be attempted first in the case of frequency distribution where only class limits are given. The mid values of the classes can be calculated by finding the average of lower and upper limits of each class and the frequency distribution will be converted as follows. Once the mid-

values are calculated they may be used for all other statistical purposes. The main assumption indirectly made is that all the values within a class are more or less having the same value as the mid-value of the class which may not be correct. However, we are not having any disadvantage because of this assumption and rather we have more advantages.

Classes (1)	Mid-Values (x_i) (2)	Frequencies f_i (3)	Total weight kg. $x_i f_i$ (4) (col. 2 × col. 3)
60.5 — 70.5	65.5	1	65.5
70.5 — 80.5	75.5	5	377.5
80.5 — 90.5	85.5	9	769.5
90.5 — 100.5	95.5	14	1337.0
100.5 — 110.5	105.5	15	1582.5
110.5 — 120.5	115.5	4	462.0
120.5 — 130.5	125.5	2	251.0
		Total 50	4845.0

$$\text{Average} = \frac{4845}{50} = 96.9 \text{ kg. per head.}$$

The above example is the frequency distribution constructed with the help of the data given in an example . The Mean calculated directly from the raw data without any grouping is 97. The difference between the Means calculated with the help of these two methods is only 0.1 (97.0 —96.9) which is very insignificant.

In this context, it may be noted that the average calculated by the first method is the correct one, though it is very laborious. When we consider the ease with which the average is calculated in the second method and also the insignificant difference, the second

method will be a more advantageous. This difference is due to the assumption as explained earlier that all the students within a particular class are having the weight equal to the mid-value of the particular class. But in actual situation this assumption is not correct. So this difference is due to the classification. As the width of the class interval is reduced, the difference in the average will also be reduced further. The lesser the class interval the lesser the difference and consequently greater the accuracy. While the average obtained from the first method is the true value of the average, the average obtained from the second method may be called as the estimate of the average. Hence the difference between the two values.

We can further simplify the computation process. Hereafter it is enough if we consider the mid-values of the class intervals only since the mid-value represents the class itself.

Short-cut Method I (Shifting the base)

As we have adopted shortcut methods for discrete distributions, we can follow similar shortcut methods.

Mid values (x)	Frequency (f)
65.5	1
75.5	5
85.5	9
95.5	14
105.5	15
115.5	4
125.5	2
Total	50

If the mid-values and the frequencies are in bigger numbers, multiplication of these two will become a problem. But we cannot alter the frequencies. However, we can reduce the size of the mid values by subtracting a convenient common arbitrary value. Here the central value namely 95.5 can be taken as the arbitrary value. The table can be rearranged as follows:

Mid value (x)	New value (d) $d = x - 95.5$	Frequency (f)	Value (df)	
(1)	(2)	(3)	(4)	
65.5	$65.5 - 95.5 = -30$	1	-30	
75.5	$75.5 - 95.5 = -20$	5	-100	-220
85.5	$85.5 - 95.5 = -10$	9	-90	
95.5	$95.5 - 95.5 = 0$	14	0	
105.5	$105.5 - 95.5 = 10$	15	150	
115.5	$115.5 - 95.5 = 20$	4	80	$+290$
125.5	$125.5 - 95.5 = 30$	2	60	
	Total	50	70	

A set of values of the new variable '*d*' are obtained. We can also find out as before the average of the new variable '*d*'. The total of all the '*d*' values is equal to 70. Therefore, the average is equal to $70 \div 50 = 1{\cdot}4$. Each of the '*d*' values is less than the corresponding '*x*' value by 95.5. kgs. Therefore, the average of '*d*' will also be less than the average of the '*x*' by 95.5 kgs.

$$d = x - 95{\cdot}5$$
$$\bar{d} = \bar{x} - 95{\cdot}5$$
$$\bar{d} + 95{\cdot}5 = \bar{x}$$
$$= 1{\cdot}4 + 95{\cdot}5 = 96{\cdot}9.$$

This is the same as the one calculated previously and this is also an estimate of the average. In the case of 'x', the values are determined from '0' or with '0' as the base, since all the values are started or counted from '0'. But in the case of 'd', values are counted from 95·5 or with the arbitrary value as the base. Here we are actually shifting the base from 0 to 95·5 (A) and converting them into new variables. Therefore, the operation involved in this process is shifting of base.

Computation of Arithmetic Mean by shifting the base and also changing the scale

In the above process involving the shifting of the base, the values of the new variable 'd' are still bigger such as 30, 20, 10 etc. They can be further reduced in their values by dividing each of the 'd' values by the class interval of the table (C) 10 in this problem. ie. 'd' values will be expressed in terms of class intervals. The working of the problem is given below:

x (1)	$y = x - 95.5$ (2)	$d = y/10$ (3)	fi (4)	dfi (5)	
65.5	−30	−3	1	− 3	
75.5	−20	−2	5	−10	−22
85.5	−10	−1	9	− 9	
95.5—A	0	0	14	0	
105.5	+10	+1	15	15	
115.5	+20	+2	4	8	+ 29
125.5	+30	+3	2	6	
		Total	50	+7	

The total value of all the 'fd' s $= 7$.

$\therefore$ Arithmetic Mean of the 'd' $= \bar{d} = 7/50 = 0{\cdot}14$.

Let $d = y/10 = y/c \;\; \bar{d} = \dfrac{\bar{Y}}{C}$

$\therefore \;\; \bar{y} = c.\ \bar{d}$

Since $y = x - 95.5$

$\bar{y} = \bar{x} - 95.5$

$\bar{y} + 95.5 = \bar{x}$

$\therefore \;\; \bar{x} = \bar{y} + 95.5$

$= 10 \times 0.14 + 95.5$ since $(\bar{x} = C\,\bar{d} + A)$

$= 1.4 + 95.5$

$=$ **96.9**

In this method the operations involved in the conversion of the 'x' variable into 'd' variable are, (1) subtraction of the arbitrary value and (2) the division by the class interval. Therefore, for calculating the Mean of 'x' from the mean of the new variable, we have to handle the operations exactly opposite to those adopted earlier and that too in the reverse direction.

First we have to multiply the mean of the new variable 'd' by the class interval 'C' and then add the arbitrary value 'A'.

Here also the Mean value obtained is only an estimate of the true Mean. This method corresponds to the conversion of Fahrenheit temperature into Centigrade and vice versa. The lowest readings in the Fahrenheit and the Centigrade thermometres are 32° and 0° respectively. In this respect, it can be said that the bases are different. The higher temparature readings in these thermometres are 212° and 100° respectively. In this, the scales are different since 180° divisions in the Fahrenheit are divided into only 100° in the centigrade.

For conversion of Fahrenheit into centrigrade, we first substract 32° from the reading and then divide the balance by

9/5 = 180/100. Hence for conversion from centigrade to Fahrenheit we first multiply the centigrade reading by 9/5 and then add 32°. It may be noted that in all the shortcut methods, the value of the Mean obtained is same.

Conversion of 113° F into Centigrade:

$$113 - 32 = 81 \div 9/5 = 81 \times \frac{5}{9} = 45°.$$

Mean of the combinations

Weighted average (or) Weighted Mean

Let us consider 3 groups of students namely A, B and C. The number of students in each group is not uniform and their average weight is also not uniform. Let us combine all the three groups and find out the average weight of the combination or the combined group. The data are given in the following table.

Group	No. of students in the group	Average weight of students (kg.)	Total weight of all the students in the group (kg.)
(1)	(2)	(3)	(4)
A	10	48	48×10 = 480
B	15	52	52×15 = 780
C	25	40	40 ×25 = 1000
Total	50		2260

$$\text{Average} = \frac{2260}{50} = 45.2 \text{ kg. per head.}$$

Though it appears somewhat cumbersome at first, it is very simple when we think over this. We have calculated the total

number of students in each group (10+15+25 = 50). Next, we have calculated the total weight of each group of students by multiplying the average weight of each group by the number of students *in the corresponding group as given in the last column* of the table. By adding the total weight of the different groups of students, we get the total weight of all the students in the combination. The final step is the total weight of all the students in the combination is divided by the total number of all the students in the combination. We get the average weight of students in the combination.

Total number *of students in the combination:* $10+15+25 = 50$.

Total weight of students in the combination

$$= (48\times10+52\times15+40\times25)$$

$$= 480+780+1000$$

$$= 2260 \text{ kg.}$$

Therefore, the average $= \dfrac{2260}{50} = 45.2$ kg.

The simple average can be calculated by adding the number of averages and dividing the total thus arrived by the number of groups.

Number of groups = 3.

The total of the averages $= 48+52+40 = 140$.

Average $= 140/3 = 46.7$ kg.

The difference between the simple average and the weighted average may be examined. These two averages will be one and the same, when the number of **students in all the groups is uniformly** equal. *In this problem,* the number *of students in* each group forms the weight of the respective group.

We can now consider the symbolic representation for this.

Group (1)	No. of students in each group (2)	Average weight of students in the group (3)	Total weight of all the students in the group (4)
A or 1	n_1	$\bar{x}_1$	$n_1\bar{x}_1$
B or 2	n_2	$\bar{x}_2$	$n_2\bar{x}_2$
C or 3	n_3	$\bar{x}_3$	$n_3\bar{x}_3$

Let $\bar{x}$ represent the average weight of the combination or the overall average weight.

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$$

Let $n_1 + n_2 + n_3 = N$. $\therefore \quad \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{N} = \bar{x}$

The above table can be re-arranged as follows:

Group	Average weight	No. of students	Total weight of all the students in the group
A	$\bar{x}_1$ (48)	n_1 (10)	$n_1\bar{x}_1 = 10 \times 48$
B	$\bar{x}_2$ (52)	n_2 (15)	$n_2\bar{x}_2 = 15 \times 52$
C	$\bar{x}_3$ (40)	n_3 (25)	$n_3\bar{x}_3 = 25 \times 40.$

In this table, the average weight of each group represented by the letter 'x' can be taken as the variable, and the number of

students of each group as the frequency of the respective group. Hence the table can again be re-written as follows.

Value x_i	Frequency f_i	$x_i f_i$
$\bar{x}_1$ (48)	f_1 (10)	$48 \times 10 = 480.$
$\bar{x}_2$ (52)	f_2 (15)	$52 \times 15 = 780.$
$\bar{x}_3$ (40)	f_3 (25)	$40 \times 25 = 1000.$
Total	50	2260

$$\bar{x} = \frac{\Sigma\, x_i f_i}{N} = \frac{\Sigma\, x_i f_i}{\Sigma f_i} = \frac{2260}{50} = 45.2 \text{ kg.}$$

Change in the formula

In the formula for the average, the numbr will be written first and value afterwards, $\frac{\Sigma\, n_i x_i}{\Sigma ni}$ whenever the number of members in each group is referred to as 'numbers'. On the other hand, the number of members in each group isr also eferred to as frequencies. Then in the formula for the average, value will be written first and the frequency afterwards.

$$\frac{\Sigma\, x_i\, f_i}{\Sigma f_i}$$

But practically there is no difference between these two formulae:

$$\bar{x} = \frac{\Sigma\, x_i f_i}{\Sigma\, f_i} = \frac{\Sigma\, n_i\, x_i}{\Sigma\, n_i}$$

It should be noted that whenever the average or Mean is mentioned, it is only the Arithmetic Mean and not any other

Mean. It may be either the simple mean or Weighted Mean depending upon the circumstances.

MEDIAN

Median occupies the second important place next to Mean in statistical analysis. Median means middle; therefore, median is difined as that value of the variable which divides the distribution into two equal halves, so that an equal number of units or individuals, or items are on either side of that Median value. Therefore, it should be clearly noted that Median is calculated with reference to the position or location of the items rather than with reference to the value of the items. The arrangement of the units or items in the order of their magnitude is very necessary for determining the value of the Median. In other words, the units have to be arranged in a frequency distribution. But such arrangement is not necessary for computing the Mean, since the values of all the items are totalled irrespective of their position in the series.

Properties of Median

1. It divides the distribution into two equal parts. Therefore, the number of units or items in each part will be equal.

2. The value of all the items in one part will be greater than the value of the Median. Similarly, the value of all the items in the other part will be less than the value of the Median.

Hence it can be termed as a locational or positional or central average. Median can be computed for ungrouped data as well as for grouped data as in the case of Mean.

Computation of Median for ungrouped data

Let us suppose the statistical details given below relate to weight of seven persons expressed in kg.

45, 39, 48, 42, 50, 35, 37.

In order to locate the Median, we have to re-arrange the value first either in the ascending or descending order of the magnitude. Let us arrange them in the ascending order as follows:

35, 37, 39, 42, 45, 48, 50.

Since there are seven items, the fourth item is the central value. Therefore, the value of the fourth item ie: 42 kg. is the median value in this series.

Let us consider another example where the number of items is an even number.

45, 35, 30, 41, 47, 38, 48, 52.

Let us rearrange the values as follows:

30, 35, 38, 41, 45, 47, 48 and 52.

As there are eight items in this series, no single item can be termed as the central item. Therefore, two items namely 4th and 5th items constitute the central items. Therefore, any one of the values (42 or 43 or 44) between 41 and 45 can satisfy the condition of the Mean. But it is always advisable to take the value in the middle of these two values 41 and 45.

$$\text{Median} = \frac{41 + 45}{2} = \frac{86}{2} = 43.$$

Median will also be expressed in the same unit of measurement as the original units.

General Formula

If there are 'n' items in the series, the $\frac{n+1}{2}$ item will be the Median value. If there are seven items in the series, $\frac{7+1}{2}$ $= 8/2. =$ 4th item will be the Median. On the other hand, if there are eight items in the series, $\frac{8+1}{2} = 9/2 = 4{\cdot}5$ th item will be the Median. In other words, it will be equal to the average value of the 4th and 5th items in the series.

Median for Discrete frequency distribution

Let us examine the following frequency distribution of marks obtained by students in a class.

Marks	No. of students
47	5
51	2
55	4
59	2
61	1
Total	14

The above 5 values (47, 51, 55, 59, 61) can be written as follows because of the frequency. The total number of items in the series will be equal to 14. The value 47 has to be repeated five times because of its frequency equal to 5. Similarly, the other values have to be repeated as many times as their respective frequency. The series will be as follows:

S.No.	(x)	**S.No. of the last item of each value**
1	47	
2	47	
3	47	
4	47	
5	47	———→ 5
6	51	
7	51	———→ 7
8	55	
9	55	
10	55	
11	55	———→11
12	59	
13	59	———→13
14	61	———→14

$$\therefore \text{ Median } = \frac{n+1}{2} \text{th item} = \frac{14+1}{2} = 7{\cdot}5$$

$\therefore$ The value of the $7\frac{1}{2}$ th item is the Median value. Values after 7th and upto 11th items are only 55.

$\therefore$ Median is 55.

In this process we have followed a laborious method of repeating the same value again and again and given the serial number. Instead of this, we can calculate the Median from the cumulative (less than) frequency.

Let us calculate the cumulative frequency of the values as follows:

Marks (x)	Frequency (f)	Cumulative frequency ($c.f.$)
47	5 ——→	5
51	2 ——→	7
55	4 ——→	11
59	2 ——→	13
61	1 ——→	14

It may be seen from this that items after 7 and upto 11 are 55. The figure noted by an arrow on the right hand side cf the value indicates nothing but the cumulative frequency which is the same as the serial number of the last item for each of the values. The same situation may arise in majority of the cases and hence the changes have been effective.

Median for continuous frequency distribution

The procedure adopted for the calculation of Median in the case of continuous frequency distribution is different from

the previous one. While we consider the cumulative frequency only in the previous case, we consider here not only the cumulative frequency but also the class limits. The cumulative frequency, the class limits, the frequency of the Median class, and the class interval of Median class are taken into consideration simultaneously.

Procedures to be followed

1. We need the true class intervals and the class limits. First we must see whether the class intervals given are true. If not we must convert them into true class intervals by suitably fitting the lower and upper limits of the classes.

2. The total frequecy should be calculated and let it be denoted by the letter N.

3. We should find half of the total frequency by dividing the total by '2' . Let it be $N/2$.

4. We should also calculate the less than cumulative frequency for each class.

5. From the cumulative frequencies, we should locate the **Median class** where the Median value or $N/2$ nd item falls.

6. We should note the lower limit of the Median class as l or L.

7. We should note the frequency of the Median class as 'f'.

8. We should note the class interval of the Median class and it should be denoted by 'c'.

9. We should calculate less than cumulative frequency of the class previous to the Median class or preceeding class of the Median class. Let it be 'm'.

10. By the method of interpolation the following formula may be adopted for calculation of the Median.

$$\text{Median} = l + \frac{(N/1 - m) \times c}{f}$$

Computation of the Median from grouped data

Median can be easily calculated from the frequency distribution. We shall consider the weight of the same 50 bags for this purpose. We can count half of the frequencies from either end of the distribution to ascertain the value of the median. For this purpose, we should consider the frequencies and the less than cumulative frequencies of the table.

Class (1)	Frequency (2)	Less than cumulative frequency (3)
60.5 — 70.5	1	1
70.5 — 80.5	5	6
80.5 — 90.5	9	15
90.5 —100.5	14	29
100.5 —110.5	15	44
110.5 —120.5	4	48
120.5 —130.5	2	50
Total	50	

The total No. of items in this problem is 50 $= N$.

$$N/2 = 50/2 = 25.$$

From the less than cumulative frequency given in col. (3), we know that there are 15 bags upto the class 80.5 — 90.5. It is understood therefore, that 15 bags are having weight less than 90.5 kg. In order to have 25 bags (N/2) we require 10 more bags (25—15 = 10) and these 10 bags have to be taken from the next class namely 90.5 — 100.5 and hence this class is known as Median class. The frequency of the Median class is 14 bags and the class interval of the Median class is 10kg. These 14 bags are arranged within a range of 10 kg. Therefore, 1 bag in this group or class is arranged within a distance of 10/14 kg.

In other words, the interval between two successive bags will be 10/14 kg. We require, 10 more bags, and these 10 bags will be arranged within a distance of $\frac{10 \times 10}{14} = 100/14 = 7.1$ kg. As this class starts with a lower limit of 90.5 kg. the 25th bag will occupy the position of $90.5 + 7.1 = 97.6$ kg. We shall apply the formula and arrive at the Median value.

Median $= M$; $N = 50$; $N/2 = 50/2 = 25$.

m = (cumulative frequency upto the previous class of the Median class). 15

$f = 14$ (frequency of the Median class).

$c = 10$ (class interval of the Median class).

$$\therefore M = l + \frac{(N/2 - m)\, c}{f}$$

$$= 90{\cdot}5 + \frac{(50/2 - 15) \times 10}{14}$$

$$= 90{\cdot}5 + \frac{(25 - 15) \times 10}{14}$$

$$= 90{\cdot}5 + \frac{10 \times 10}{14}$$

$$= 90{\cdot}5 + 7{\cdot}1 = 97{\cdot}6 \text{ kg.}$$

Type II

Let us calculate the Median from the following:

Fortnightly wages (Rs.) (1)	No. of workers (2)	Cumulative frquencies (3)
21 – 30	2	2
31 — 40	5	7
41 — 50	12	19
51 — 60	9	28
61 — 70	4	32
71 — 80	2	34

There is a difference between the previous type and the present type. In the previous type, the class intervals are true class intervals, whereas in this case, the class intervals are not true. Therefore, the class intervals have to be converted into true class intervals as follows:

Fortnightly wages Rs. (1)	No. of workers (2)	Cumulative frequency (3)
20.5 — 30.5	2	2
30.5 — 40.5	5	7
40.5 — 50.5	12	19
50.5 — 60.5	9	28
60.5 — 70.5	4	32
70.5 — 80.5	2	34
	34	

$N = 34;\ N/2 = 34/2 = 17.$

Median class $= 40.5 - 50.5$

$$\begin{aligned} c &= 30.5 - 40.5 = 10 \\ l &= 40.5 \\ f &= 12. \\ m &= 7 \end{aligned}$$

$$\begin{aligned} M &= l + \frac{(N/2 - m) \times c}{f} \\ &= 40.5 + \frac{(17 - 7)}{12} \times 10 \\ &= 40.5 + 10/12 \times 10 \\ &= 40.5 + \frac{100}{12} = 40.5 + 8.3 = \text{Rs. } 48.8. \end{aligned}$$

Type III

Mid value (1)	Frequency (2)
15	7
25	6
35	9
45	4
55	5
65	4
75	7
85	4
	46

In this example the mid-values of the classes are given instead of the class limits. The limits of various classes have to be determined from the mid-values of the classes. In order to fix the class limits, we require class intervals. The interval between two successive mid-values will be the class interval. In this particular case, the difference between successive midvalues is 10 and hence 10 can be taken as the class interval. Hence half the value of class-interval can be subtracted from the midvalue to find out the lower limit and half the value of the class interval can be added to the mid-value to find out the upper limit. Thus the true class intervals can be calculated as follows:

Class interval (1)	Frequency (2)	Cumulative frequency (3)
10 — 20	7	7
20 — 30	6	13
30 — 40	9	22
40 — 50	4	26
50 — 60	5	31
60 — 70	4	35
70 — 80	7	42
80 — 90	4	46
	46	

$N = 46$. $N/2 = 23$; $m = 22$. Median class $= 40 - 50$.

$l = 40$; $c = 10$; $f = 4$.

$$M = l + \frac{(N/2 - m) \times c}{f}$$

$$= 40 + \frac{(23 - 22)\, 10}{4}$$

$$= \mathbf{40 + 2.5 = 42.5}$$

Type IV

Calculation of Median when cumulative frequencies without frequencies are given

Let us consider the following example.

Marks less than	Cumulative frequency
10	3
20	5
30	8
40	12
50	20
60	25

This problem is slightly different from the previous one. In this problem, upper class limits of the classes are given. Therefore, we have to find out the lower class limits of each class and establish each class. Since we are given the cumulative frequency, we should calculate the actual frequency by subtracting the

cumulative frequency of the preceeding class from the cumulative frequency of the class and construct the table:

Class	Cumulative frequency	Frequency
0 — 10	3	3 — 0 = 3
10 — 20	5	5 — 3 = 2
20 — 30	8	8 — 5 = 3
30 — 40	12	12 — 8 = 4
40 — 50	20	20 — 12 = 8
50 — 60	25	25 — 20 = 5
		25

The procedures for calculation of Median afterwards are same as before.

$N = 25; \; N/2 = 25/2 = 12.5$

Median class $= 40 - 50$. $\quad l = 40; \quad m = 12; f = 8; \; c = 10.$

$$M = l + \frac{(N/2 - m)\,c}{f} \qquad = 40 + \frac{(25/2 - 12) \times 10}{8}$$

$$= 40 + \tfrac{1}{2} \times \frac{10}{8}$$

$$= \mathbf{40.6}$$

Computation of Median from graphs

We can also calculate the Median from the graph. As we require less than cumulative frequency, we can compute the Median from the graph for the less than cumulative frequency. In the graph, the values of the frequencies will be plotted on the Y-axis.

As the Median value relates to the middle portion of the distribution ($N/2$), we shall locate the position on the Y-axis equivalent to $N/2$. From this point on the Y-axis, a line parallel to the X-axis can be drawn to cut the ogive curve. From this point of

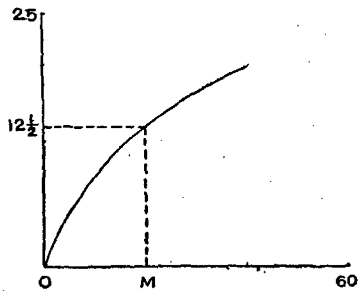


FIG. 12
Median

intersection on the ogive curve, a line parallel to the Y-axis and perpendicular to X-axis can be drawn, cutting X-axis. The point of intersection on the X-axis will indicate Median value.

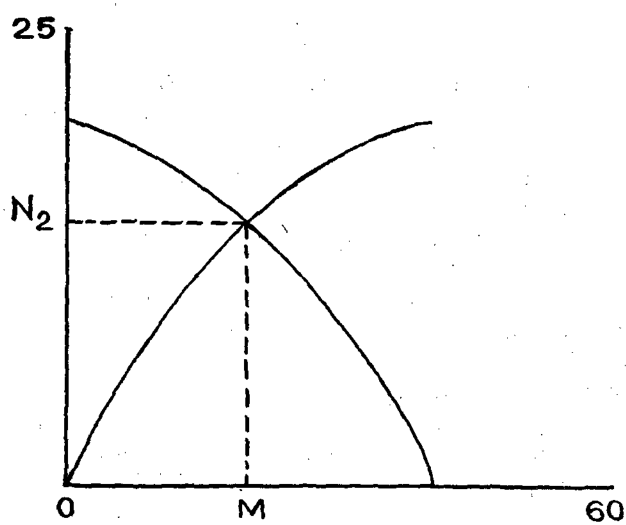


FIG. 13
Median

We can use the curve for either lower than cumulative frequency or greater than cumulative frequency. In both the cases, the procedures are same. However, we can use both the curves simultaneously. From the point of intersection of these two curves for less than and greater than cumulative frequencies, we should draw a perpendicular to X-axis. Irrespective of the fact whether we use less than or greater than cumulative frequency curve, the value of $N/2$, is same and consequently the line drawn from the point on Y axis corresponding to $N/2$, will cut the ogive curve on the same point. The point of intersection of the perpendicular on the X-axis will be the Median value.

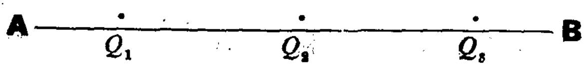
Property of the median

1. An important property of the Median is that it is inflenced by the position of the items in the array and not by the value or size of the items.
2. The median will be expressed in the same unit of measurement as the original items in the distribution.

QUARTILE (Q)

We have seen that the Median divides the number of units in the distribution into two equal halves. There are other measures similar to Median which divide the distribution into many parts. We can divide the distribution into four equal parts or five equal parts or 10 equal parts or 100 equal parts. For each kind of division we have different kinds of measures and they are also called measures of central tendency.

Quartiles are another set of measures of central tendency which divide the distribution into 4 equal parts. In dividing the distribution into four equal parts, we should have three points of location and the values of these dividing points are called Quartiles. Since there are three dividing points, they are designated with the order of their arrangement or occurrence namely First Quartile (Q_1), Second Quartile (Q_2) and Third Quartile (Q_3).



First Quartile or Lower Quartile

The first quartile divides the distribution into two unequal parts such that 1/4th of the items or 25% of the units or individuals will have values less than the value of the first quartile value and the remaining 3/4th or 75% of the units or individuals or classes will have values greater than the value of the first quartile.

I. Calculation of the lower quartile from ungrouped data

1. The given items or values should be arranged in the ascending order of their magnitude.

2. The total number of items may be noted by the letter 'n'.

The lower quartile (Q_1) should represent the value of $\frac{(n+1)}{4}$ th item.

If there are seven values as follows:

40, 25, 35, 50, 60, 20, 75, they have to be re-arranged as follows in the ascending order of the magnitude.

20, 25, 35, 40, 50, 60, 75.

$n = 7; \quad n+1 = 7+1 = 8.$

The second item represents the Q_1. Therefore, the value of the second item ie. 25 represents the Q_1.

In case the $(n+1)$ is not exactly divisible by 4, the following procedures may be adopted.

Suppose there are 10 items in an example:

20, 25, 35, 40, 60, 70, 72, 79, 80, 85.

$n = 10; \quad n+1 = 11; \quad (n+1)/4 = 11/4 = 2.75 \ (2\frac{3}{4}).$

The lower quartile lies between the second and the third items.

Value of the second item $=$ 25.

Value of the third item $=$ 35

Difference between the second and third values $= 35 - 25 = 10.$

$\frac{3}{4}$th *difference between second* and the third items is $= \dfrac{10 \times 3}{4} = 7.5$

$\therefore \quad Q_1$ = Value of second item + 3/4 th difference between second and third items.

$= 25 + 7.5 = \mathbf{32.5}$

If there are eight items in the series, the value of the $(8+1)/4$ th ie: 2.25th item will be the lower quartile. In other words, it will be equal to the sum of the values of the second item and the $\frac{1}{4}$th *value of the difference between the values of the second* and third items.

If there are 9 items in the series, the value of $(9+1)/4 = 2.5$th item will be the lower quartile. This is equal to the sum of the value of the second item and half the difference between the value of the second and third items.

II. Calculation of lower quartile from discrete frequency distribution

The following procedures may be adopted:

1. The given frequency distribution should be converted into less than cumulative frequency.

2. The total sum of all the frequencies will be denoted by 'N'.

3. The value of $\dfrac{N+1}{4}$ is found out and this will represent the lower quartile.

Value of the items	Frequency	Cumulative frequency
25	3	3
35	4	7
45	2	9
55	5	14
	14	

$N = 14 \qquad (N+1)/4 = 15/4 = 3.75$

$\therefore \quad Q_1 =$ Value of $3\frac{3}{4}$ th item.

All the items beyond third and upto 7th item are having values equal to 35.

$\therefore \quad 3\frac{3}{4}$th item is 35.

III. Calculation of lower quartile from continuous frequency distribution

1. The frequencies should be converted into less than cumulative frequencies.

2. The total of all the frequencies should be calculated (N) and this should be divided by 4 ($N/4$) to find out $\frac{1}{4}$th of the total frequency.

3. We should find out the lower quartile class in which Q_1 lies.

4. The true lower limit of the lower quartile class should be determined (l).

5. The frequency of the lower quartile class should also be noted (f).

6. The less than cumulative frequency of the class preceeding the lower quartile class should be noted (m).

7. The class interval of the lower quartile class should be noted(c).

Q_1 = Lower quartile;

l = lower limit of the lower quartile class.

f = frequency of the lower quartile class.

m = cumulative frequency of the class previous to the first quartile class.

c = The class interval of the first quartile class.

The following formula can be used for calculating the lower quartile:

$$Q_1 = l + \frac{(N/4 - m)\,c}{f}$$

Class	Frequency	Cumulative frequency
0 — 5	3	3
5 — 10	4	7
10 — 15	2	9
15 — 20	5	14
20 — 25	1	15

$N = 15$; $N/4 = 15/4 = 3.75$

First quartile class = 5 — 10.

$l = 5;\ f = 4;\ m = 3;\ c = 5.$

$$Q_1 = 5 + \frac{(15/4 - 3) \times 5}{4}$$

$$= 5 + \frac{0.75 \times 5}{4}$$

$$= 5 + 0.94 = 5.94.$$

IV. Calculation from the graph

As in the case of Median, the Quartiles can be compiled from the Ogive curve. We should draw the Ogive curve for less than the cumulative frequency. The values are plotted on X-axis and the cumulative frequencies are plotted on the 'Y' axis. As the first quartile (Q_1) or Lower quartile divides the distribution in the ratio 1:3, we should select the point on the Y-axis corresponding to the frequency $N/4$. From this point on the Y-axis, a straight line parallel to the X-axis should be drawn to cut the

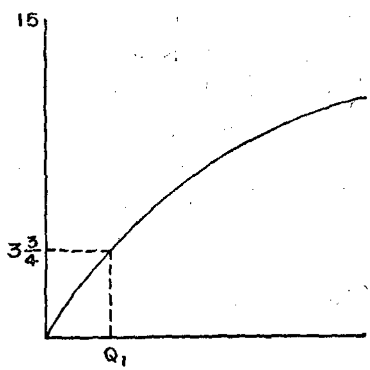


FIG. 14
First quartile

Ogive curve. From this point of intersection on the Ogive curve, a straight line perpendicular to the X-axis or parallel to Y-axis should be drawn to cut the X-axis. This point of intersection on the X-axis indicates location of the first quartile. Consequently, the value of this point on the X-axis will give the value of the first quartile.

Property

1. The first quartile will be expressed in the same unit of measurement as the original items in the distribution.
2. The first quartile is influenced by the position of the items in the array and not by the value or size of the items.

Third Quartile (Q_3)—Upper Quartile

The upper quartile divides the distribution into two unequal parts so that one fourth (1/4) of the items will have value greater than Q_3 and three fourths (3/4) of the items will have value less than Q_3. In fact, it is quite opposite to Q_1. This can be computed in the same way as the first quartile.

I. Calculation of upper quartile from ungrouped data

The following methods are adopted:

(*i*) The given items should be arranged in the ascending order of their magnitude.

(*ii*) The total number of items should be found out and it should be denoted by '*n*'.

(*iii*) The following formula may be used:

$$Q_3 = \text{Value of } 3(n+1)/4\text{th item.}$$

$$\text{(or) } 3/4\ (n+1)\text{th item.}$$

Let us calculate the upper quartile:

29, 35, 26, 44, 40, 55.

Let us re-arrange them in the ascending order.

26, 29, 35, 40, 44, 55.

The total number of items (n) = 6.

Q_3 = 3/4 (n+1) th item.

$$= \frac{3\,(n+1)}{4} = 3/4\,(6+1) = \frac{3 \times 7}{4} = \frac{21}{4} = 5.25.$$

We have to find out the value of the $5\frac{1}{4}$ item.

Q_3 = Value of the 5th item + $\frac{1}{4}$ (Value of the 6th item — value of the 5th item).

Value of the 5th item = 44.

Value of the 6th item = 55.

Difference between the 6th and 5th items = 55—44 = 11.

$\frac{1}{4}$th difference = 11/4 = 2.75

$\therefore$ Q_3 = Value of the 5th item + $\frac{1}{4}$th difference

= 44 + 2.75

= **46.75**

(1) Calculate the upper quartile for the following data

29, 35, 26, 40, 58, 44, 55.

First re-arrange them in the ascending order.

26, 29, 35, 40, 44, 55, 58.

$n = 7$. Q_3 = value of $\frac{3(n+1)}{4}$ th item.

$= \frac{3(7+1)}{4} = \frac{3 \times 8}{4}$ = 6th item.

Value of the 6th item = 55.

$\therefore$ $Q_3 = 55$.

II. Calculation of upper quartile from discrete frequency distribution

1. For calculation of upper quartile for discrete frequency distribution, we should convert the frequency into less than cumulative frequency.

2. The total sum of all the frequencies may be denoted by the letter 'N'.

3. The same formula i.e: $3(N+1)/4$ can be used.

We shall calculate the third quartile for the following distribution:

Value	Frequency
25	3
35	4
45	2
55	5
	14

We should first calculate the less than the cumulative frequency

Value (1)	Frequency (2)	Cumulative frequency (3)
25	3	3
35	4	7
45	2	9
55	5	14

Total frequency $N = 14$.

$\therefore \quad N+1 = 14+1=15.$

$\frac{3}{4}(N+1) = \frac{3}{4} \times 15 = 45/4 = 11.25$

It is seen from col(3) of the above table that all the items after 9 and upto 14 are having values 55. As the 11¼ (11.25)th item lies after 9, Q_3 is equal to 55.

III. Calculation of upper quartile from continuous frequency distribution

In the case of continuous frequency distribution, the formula adopted is slightly different from the formula adopted for discrete frequency distribution. However, the procedures are same. The formula adopted is $\frac{3N}{4}$ instead of $\frac{3(N+1)}{4}$

Methods followed

1. The frequencies should be converted into less than cumulative frequency.

2. The total of the frequencies should be found out and indicated by the letter 'N'.

3. It should be divided by 4 and multiplied by 3, or it should be multiplied by 3 and then divided by 4. In other words, it should be multiplied by $\frac{3}{4} = 3N/4$.

4. We should find out the class in which the Q_3 lies and this class is known as upper quartile or third quartile class.

5. The true lower limit of the upper quartile class should be ascertained (l).

6. The frequency of the upper quartile class should be noted (f).

7. The less than cumulative frequency of the class preceeding the upper quartile class should be noted (m).

8. The class interval of the upper quartile class should be noted (c).

Upper quartile: $Q_3 = l + \frac{(3N/4 - m)}{f} \times C$

Q_3 = Upper quartile; l = Lower limit of the upper quartile class.

N = total frequency. m = Less than cumulative frequency of the class previous to the upper quartile class.

c = Class interval of the upper quartile class.

f = frequency of the upper quartile class.

Calculate the upper quartile for the following distribution:

Class (1)	Frequency (2)	Less than cumulative frequency (3)
0 — 5	3	3
5 — 10	4	7
10 — 15	2	9
15 — 20	5	14
20 — 25	1	15
Total	15	

Total frequency $N = 15$. $3N/4 = \frac{3 \times 15}{4} = 45/4$

$= 11.25$ ($11\frac{1}{4}$)th item.

Q_3 = Value of the item 11.25.

We know from col. (3) of the above table that $11\frac{1}{4}$th item lies in the class 15 to 20 and this is the upper quartile class.

Upper quartile class = 15 — 20

The true lower limit of the upper quartile class (l) = 15.

Frequency of the upper quartile class (f) = 5.

Class interval of the upper quartile class (c) = 20 — 15 = 5.

The less than cumulative frequency of the class preceeding the upper quartile class. = 9.

$$\therefore Q_3 = l + \frac{(3N/4 - m) \times c}{f}$$

$$= 15 + \frac{(45/4 - 9) \times 5}{5}$$

$$= 15 + \frac{(11.25 - 9) \times 5}{5}$$

$$= 15 + \frac{2.25 \times 5}{5}$$

$$= 15 + 2.25$$

$$= 17.25$$

Upper quartile = 17.25

IV. Calculation of upper quartile from the graph

As in the case of Median and Lower quartile the upper quartile can also be computed from the Ogive curve. We should draw Ogive curve for less than cumulative frequency distribution. The values corresponding to the lower limits of the various classes are plotted on the X-axis and the cumulative frequencies are

plotted on the Y-axis. As the third quartile devides the distribution in the ratio 3:1, we should select the point of location on the *Y*-axis corresponding to $3N/4$. From this point on the *Y*-axis, we should draw a straight line parallel to the *X*-axis cutting the Ogive curve. From this point of intersection on the Ogive

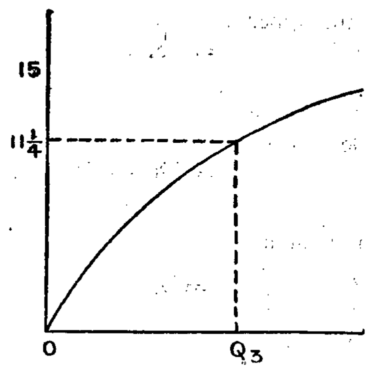


FIG. 15
Third quartile

curve, we should draw a straight line perpendicular to the *X*-axis or parallel to the *Y*-axis to cut the *X*-axis. The point of intersection on the *X*-axis will correspond to the value of Q_8 item. Therefore, the value of *X* corresponding to this point of intersection on the *X*-axis relates to the upper quartile or third quartile.

Properties

1. The third quartile will be expressed in the same unit of measurement as the original values in the distribution.

2. The third quartile is influenced by the position of the items in the array and not by the value or size of the item.

We have referred to the lower quartile as the first quartile Q_1 and the upper quartile as the third quartile (Q_8). Naturally, we may be tempted to ask about the missing quartile namely the second quartile. The second quartile is nothing but the Median (Q_2).

QUINTILES

We have stated earlier that a distribution can be divided into four equal parts by the first, second and third quartiles. In the same way the given distribution can be divided into 5 equal parts by the various Quintiles, namely, First, Second, Third, Fourth quintiles.

The various quintiles can be calculated for a given series of data as in the case of quartiles. Therefore, the procedures are more or less same, the points of location may differ and consequently the formulae have slightly to be altered.

For ungrouped data and discrete frequency distribution the following formulae have to be adopted:

$$\text{First Quintile} = \frac{(n+1)}{5}$$

$$\text{Second Quintile} = \frac{2\,(n+1)}{5}$$

$$\text{Third Quintile} = \frac{3\,(n+1)}{5}$$

$$\text{Fourth Quintile} = \frac{4\,(n+1)}{5}$$

For continuous frequency distribution, the following formulae can be followed:

$$\text{First Quintile} = l + \frac{(N/5 - m)\,c}{f}$$

$$\text{Second Quintile} = l + \frac{(2N/5 - m)\,c}{f}$$

$$\text{Third Quintile} = l + \frac{(3N/5 - m)\,c}{f}$$

$$\text{Fourth Quintile} = l + \frac{(4N/5 - m)\,c}{f}$$

These different quintiles can be calculated from the graph also, by drawing straight lines parallel to X-axis to cut the less than cumulative frequency curve. But these straight lines should be drawn from the parts on the Y-axis, corresponding to the frequency namely, $N/5$, $2N/5$, $3N/5$ and $4N/5$ for the first, second, third and fourth quintiles respectively.

Properties

1. The quintiles will be expressed in the same unit of measurement as the original items in the distribution.
2. They are influenced by the position of the items and not by the value or size of the items.

DECILES

We can divide the distribution into 10 equal parts. In this process we get 9 dividing points or positions and they are called Deciles.

These nine deciles are generally denoted by the letters D_1, D_2, D_3 D_4, D_5, D_6, D_7, D_8 and D_9.

Calculation of deciles for ungrouped data

The given items should first be arranged in the order of their magnitude. The total number of items in the array should be indicated by the letter 'n'. Then $(n+1)$th item should also be found out. Then this should be divided by 10 = $(n+1)/10$. This quotient obtained should be multiplied by the number of respective deciles as follows: After these, the respective deciles can be calculated with the help of the formula noted against each.

D_1 = Value of $(n+1)/10$th item;

D_2 = Value of $2(n+1)/10$th item;

D_3 = Value of $3(n+1)/10$th item;

D_4 = Value of $4(n+1)/10$th item;

D_5 = Value of $5(n+1)/10$ th item;

D_6 = Value of $6(n+1)/10$ th item;

D_7 = Value cf $7(n+1)/10$ th item;

D_8 = Value of $8(n+1)/10$ th item;

D_9 = Value of $9(n+1)/10$ th item;

Let us consider the following items and work out the various deciles:

26, 15, 20, 18, 25, 17, 29, 40, 35, 28, 31, 42, 50,
53, 60, 65.

These values should first be arranged in the order of their magnitude as follows:

15, 17, 18, 20, 25, 26, 28, 29, 31, 35, 40, 42. 50,
53, 60, 65.

The total number of items = 16.

$$\therefore \quad n + 1 = 17.$$

$$(n + 1)/10 = 17/10$$

$$D_1 = \frac{1(n+1)}{10} = \frac{17}{10} = 1.7\text{th item}$$

$$D_2 = \frac{2(n+1)}{10} = \frac{2 \times 17}{10} = 3.4\text{th item.}$$

$$D_3 = \frac{3\ (n+1)}{10} = \frac{3 \times 17}{10} = 5.1\text{th item.}$$

$$D_4 = \frac{4\ (n+1)}{10} = \frac{4 \times 17}{10} = 6.8\text{th item}$$

$$D_5 = \frac{5\ (n+1)}{10} = \frac{5 \times 17}{10} = 8.5\text{th item}$$

$$D_6 = \frac{6\ (n+1)}{10} = \frac{6 \times 17}{10} = 10.2\text{th item.}$$

$$D_7 = \frac{7\ (n+1)}{10} = \frac{7 \times 17}{10} = 11.9\text{th item.}$$

$$D_8 = \frac{8\ (n+1)}{10} = \frac{8 \times 17}{10} = 13.6\text{th item.}$$

$$D_9 = \frac{9\ (n+1)}{10} = \frac{9 \times 17}{10} = 15.3\text{th item.}$$

D_1 = 1.7th item = First item + 0.7 × difference between first and second items.

$$= 15 + 0.7\ (17-15)$$

$$= 15 + 0.7 \times 2$$

$$= 15 + 1.4 = 16.4$$

D_2 = 3.4th item = 3rd item + 0.4 × difference between 3rd and 4th items.

$$= 18 + 0.4\ (20 - 18)$$

$$= 18 + 0.8$$

$$= 18.8$$

D_3 = 5.1th item = 5th item + 0.1 × difference between 5th and 6th items.

$$= 25 + 0.1\ (26 - 25)$$

$$= 25 + 0.1$$

$$= 25.1$$

D_4 = 6.8th item = 6th item + 0.8 × difference between 6th and 7th items.

$$= 26 + 0.8\ (28 - 26)$$

$$= 26. + 1.6$$

$$= 27.6$$

D_5 = 8.5th item = 8th item + 0.5 × difference between 8th & 9th items.

$$= 29 + 0.5 \quad (31 - 29)$$

$$= 29 + 1.0$$

$$= 30$$

D_6 = 10.2th item = 10th item + 0.2 × difference between 10th & 11th items.

$$= 35 + 0.2 \times (40 - 35)$$

$$= 35 + 1.0$$

$$= 36$$

D_7 = 11.9th item = 11th item + 0.9 × difference between 11th and 12th items.

$$= 40 + 0.9 \quad (42 - 40)$$

$$= 40 + 1.8$$

$$= \mathbf{41.8}$$

D_8 = 13.6th item = 13th item + 0.6 × difference between 13th & 14th items.

$$= 50 + 0.6 \quad (53 - 50)$$

$$= 50 + 1.8$$

$$= \mathbf{51.8}$$

D_9 = 15.3th item = 15th item + 0.3 × difference between 15th and 16th items.

$$= 60 + 0.3 \ (65 - 60)$$

$$= 60 + 1.5$$

$$= \mathbf{61.5}$$

Type II

Calculation of deciles for discrete frequency distribution:

1. In the case of discrete frequency distribution, the less than cumulative frequency should first be calculated.

2. The total number of items or the total of all the frequencies should be indicated by N.

Afterwards the deciles are calculated with the help of the following formulae:

D_1 = Value of 1 $(N+1)/10$th item;

D_2 = Value of $2(N+1)/10$th item;

D_3 = Value of $3(N+1)/10$th item;

D_4 = Value of $4(N+1)/10$th item:

D_5 = Value of $5(N+1)/10$th item;

D_6 = Value of $6(N+1)/10$th item;

D_7 = Value of $7(N+1)/10$th item;

D_8 = Value of $8(N+1)/10$th item;

D_9 = Value of $9(N+1)/10$th item.

We shall consider the following example:

Value	Frequency	Cumulative frequency
25	4	4
35	3	7
45	6	13
55	7	20
65	3	23
75	1	24

$N = 24; \quad N+1 = 24+1 = 25.$

$$N+1/10 = 25/10 = 2{\cdot}5$$

$$D_1 = 1\,(N+1)/10 = 1(24+1)/10\text{th item}$$
$$= 25/10\text{th item}$$
$$= 2{\cdot}5\text{th item.}$$

The value of 2.5th item = 25.

$$D_2 = 2\,(N+1)/10 = \frac{2 \times 25}{10} = 5\text{th item.}$$

Value of 5th item = 35.

$$D_3 = 3\,(N+1)/10 = \frac{3 \times 25}{10}\text{ th item}$$
$$= 7.5\text{th item.}$$

Value of the 7.5th item = 45.

$$D_4 = 4\,(N+1)/10 = \frac{4 \times 25}{10}\text{ th item}$$
$$= 10\text{th item.}$$

Value of the 10th item = 45.

$$D_5 = 5\,(N+1)/10 = \frac{5 \times 25}{10} = 12.5\text{th item}$$

Value of the 12.5th item = **45**

$$D_6 = 6\,(N+1)/10 = \frac{6 \times 25}{10}\text{ th item}$$
$$= 15\text{th item.}$$

Value of the 15th item = 55.

$$D_7 = 7\,(N+1)/10 = \frac{7 \times 25}{10}\text{ th item.}$$
$$= 17.5\text{ th item}$$

Value of the 17.5th item $= 55.$

$$D_8 = 8\,(N+1)/10 = \frac{8 \times 25}{10} = 20\text{th item.}$$

Value of the 20th item $= 55.$

$$D_9 = \frac{9(N+1)}{10} = \frac{9 \times 25}{10}\text{ th item}$$

$$= 22.5\text{th item.}$$

Value of 22.5th item $= 65.$

III. Calculation of deciles from continuous frequency distribution

We can also calculate deciles for continous frequncy distribution as we have calculated Median, Quartiles and Quintiles. The only change introduced in the formula is substitution of $N/10$ for $N/2$ in the case of median, for $N/4$ in the case of first quartile, for $3N/4$ in the case of third quartile etc.

The procedures adopted are as follows:

1. The cumulative frequencies should be calculated.
2. The total frequency should be denoted by the letter 'N'.
3. The true lower limit of the class should also be fixed.

The formulae will emerge as follows:

$$D_1 = l + \frac{(N/10 - m)\,c}{f}$$

$$D_2 = l + \frac{(2N/10 - m)\,c}{f}$$

$$D_3 = l + \frac{(3N/10 - m)\,c}{f}$$

$$D_4 = l + \frac{(4N/10 - m)\,c}{f}$$

$$D_5 = l + \frac{(5N/10 - m)\,c}{f}$$

$$D_6 = l + \frac{(6N/10 - m)\,c}{f}$$

$$D_7 = l + \frac{(7N/10 - m)\,c}{f}$$

$$D_8 = l + \frac{(8N/10 - m)\,c}{f}$$

$$D_9 = l + \frac{(9N/10 - m)\,c}{f}$$

In these formulae:

l = lower limit of the respective decile class.

c = class interval of the respective decile class.

f = frequency of the respective decile class.

m = cumulative frequency of the class preceeding the respective decile class.

We shall examine this in detail with the help of an example:

Class (1)	Frequency (2)	Cumulative frequency (3)
0 — 5	3	3
5 — 10	4	7
10 — 15	7	14
15 — 20	8	22
20 — 25	2	24
25 — 30	1	25
Total	25	

From the frequencies given in col (2) we should calculate the cumulative frequency as given in col (3) of the table.

Total frequency (N) = 25.

D_1 relates to $N/10$th item. $\therefore$ $N/10 = 25/10 = 2.5$

$$D_1 = l + \frac{(N/10 - m)\,c}{f}$$

$$= 0 + \frac{(25/10 - 0) \times 5}{3}$$

$$= \frac{25}{10} \times \frac{5}{3} = 4.17$$

D_2 relates to $2N/10$th item.

$$= 2N/10 = \frac{2 \times 25}{10} = 5\text{th item.}$$

5th item lies in the class 5 — 10.

$$D_2 = l + \frac{(2N/10 - m)\,c}{f}$$

$$= 5 + \frac{\left(\frac{2 \times 25}{10} - 3\right) \times 5}{4}$$

$$= 5 + \frac{(5 - 3) \times 5}{4}$$

$$= 5 + 2.5 = \mathbf{7.5}$$

$$D_3 = 3N/10 = \frac{3 \times 25}{10} = 7.5$$

D_3 class = (10 – 15)

$$= 10 + \frac{(7.5 - 7) \times 5}{7}$$

$$= 10 + \tfrac{1}{2} \times 5/7 = 10\ \frac{5}{14}$$

$$D_4 = 4N/10 = \frac{4 \times 25}{10} = 10.$$

D_4 class $= 10 - 15$

$$= 10 + \frac{(10 - 7) \times 5}{7}$$

$$= 10 + \frac{3 \times 5}{7} = 12\tfrac{1}{7}.$$

$$D_5 = 5N/10 = \frac{5 \times 25}{10} = 12.5$$

D_5 Decile class $= 10 - 15.$

$$D_5 = 10 + \frac{(12.5 - 7) \times 5}{7}$$

$$= 10 + \frac{5.5 \times 5}{7}$$

$$= 10 + \frac{27.5}{7}$$

$$= 10 + 3.93 = 13.93$$

$$D_6 = 6N/10 = \frac{6 \times 25}{10} = 15.$$

Decile class $= 15 - 20.$

$$D_6 = 15 + \frac{(15 - 14) \times 5}{8}$$

$$= 15 + 5/8 = 15.63$$

$$D_7 = 7N/10 = \frac{7 \times 25}{10} = 17.5$$

Decile class $= 15 - 20.$

$$D_7 = 15 + \frac{(17.5 - 14) \times 5}{8}$$

$$= 15 + \frac{3.5 \times 5}{8} = 15 + \frac{17.5}{8}$$

$$= 17.2$$

$$D_8 = 8N/10 = \frac{8 \times 25}{10} = 20.$$

Decile class $= 15 - 20$

$$D_8 = 15 + \frac{(20 - 14) \times 5}{8}$$

$$= 15 + \frac{6 \times 5}{8} = 18.75$$

$$D_9 = 9N/10 = \frac{9 \times 25}{10} = 22.5$$

Decile class $= 20 - 25.$

$$\therefore D_9 = 20 + \frac{(22.5 - 22) \times 5}{2}$$

$$= 20 + \frac{(0.5 \times 5)}{2}$$

$$= 21.25$$

Calculation from the graph

Deciles can be calculated from the Ogive curve for the less than cumulative frequency distribution. Corresponding to the items $N/10$, $2N/10$, $3N/10$, $4N/10$, $5N/10$, $6N/10$, $7N/10$, $8N/10$ and $9N/10$, points on the Y-axis should be located. From these points, we should draw straight lines parallel to X-axis to cut the Ogive curve at different points called p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8 and p_9. From these points on the Ogive curve, we should draw straight lines perpendicular to X-axis to cut the X-axis at

9 different points denoted by the letters x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 and x_9. The values of X-axis at these 9 points will indicate the respective decile value.

PERCENTILES

Another measure of location of central tendency is Percentile. These points of location divide the distribution or series into 100 equal parts and as such there are 99 points of location. They are called Percentiles and denoted by the letter 'P' with the suffix corresponding to the serial number of the percentile (eg.): p_1, p_2, p_3,....p_{25}..... p_{99}.

1. Calculation of Percentiles from grouped data

The methods are similar to those adopted in the case of quartiles, quintiles and deciles. The only difference is that the value of $(n+1)$ should be divided by 100. The values of the different percentiles would relate to the value of the items mentioned against each given below:

P_1 = Value of $(n+1)/100$th item

P_2 = Value of $2(n+1)/100$,, ,,

P_3 = Value of $3(n+1)/100$,, ,,

P_5 = Value of $5(n+1)/100$,, ,,

P_{10} = Value of $10(n+1)/100$,, ,,

P_{25} = Value of $25(n+1)/100$,, ,,

P_{50} = Value of $50(n+1)/100$,, ,,

P_{99} = Value of $99(n+1)/100$,, ,,

The general formula can be written as follows:

$$P_r = \text{Value of } \frac{r(n+1)}{100}\text{th item}$$

Where 'r' can take any value between 1 and 99. n, represents the total number of items in the series. Let us examine the following item and calculate 15th, 30th and 65th percentiles.

21, 28, 34, 15, 30, 70, 65, 48, 51, 75.

Let us re-arrange these values in the ascending order as follows:

15, 21, 28, 30, 34, 48, 51, 65, 70, 75;

The total number of items $= n = 10$.

$\therefore \; n + 1 = 10 + 1 = 11.$

(1) $P_{15} = \dfrac{15 \times (10 + 1)}{100}$ th item.

$= \dfrac{15 \times 11}{100} = 165/100 = 1.65$th item.

Value of the 1.65th item = Value of the 1st item + 0.65 × difference of the 1st and 2nd items.

$= 15 + 0.65\ (21 - 15)$

$= 15 + 0.65 \times 6$

$= 15 + 3.90$

$= 18.90$

(2) $P_{30} = \dfrac{30 \times (10 + 1)}{100}$ th item.

$= \dfrac{30 \times 11}{100} = 3.3$th item

Value of 3.3th item = Value of 3rd item + 0.3 × difference between 3rd & 4th items.

$= 28 + 0.3\ (30 - 28)$

$= 28 + 0.3 \times 2$

$= 28.6$

(3) $P_{65} = 65\,(10 + 1)/100 + 7.15$

The value of the 7.15th item $=$ Value of the 7th item $+ 0.15 \times$ difference between 7th&8th items.

$= 51 + 0.15\,(65 - 51)$

$= 51 + 2.10$

$= 53.10$

II. Calculation of percentiles from discrete frequency distribution

The given frequencies should be converted into less than cumulative frequencies. Sum of the total frequencies should be calculted.

The percentiles should be calculated as follows:

P_1 = Value of the N+1/100th item.

P_{25} = Value of the 25(N+1)/100th item.

P_{99} = Value of the 99(N+1)/100th item.

Example

Value of the item	Frequency	Cumulative frequency
25	4	4
35	6	10
45	5	15
55	7	22
65	3	25
75	2	27
85	2	29

$N = 29. \quad N+1 = 30.$

$\therefore N + 1/100 = 30/100 = 3/10$

Let us calculate first the less than cumulative frequency as given in col. 3.

$$P_5 = \frac{5\,(29+1)}{100} = \frac{5 \times 30}{100} = 1.5\text{th item}$$

Value of the 1.5 th item = 25.

$$P_{20} = \frac{20\,(29+1)}{100} = \frac{20 \times 30}{100} = 6\text{th item.}$$

Value of the 6th item = 35.

$$P_{44} = \frac{44\,(29+1)}{100} = \frac{44 \times 30}{100} = 13.2\text{th item.}$$

Value of the $13\frac{1}{5}$th item = 45.

III. Calculation of Percentiles from continuous frequency distribution

The true lower limits of the classes should be calculated.

The sum of the total frequencies should be arrived at. The less than cumulative frequency should be calculated.

Let us denote P_r = N/100th item.

Let us calculate P_{15}, P_{22}, P_{70}th from the following table.

Class	Frequency	Cumulative frequency
0 — 5	4	4
5 — 10	6	10
10 — 15	5	15
15 — 20	7	22
20 — 25	3	25
25 — 30	2	27
30 — 35	3	30
Total	30	

$N = 30; \quad \therefore \; N/100 = 30/100 = 3/10.$

$$P_{15} = \frac{15 \times 30}{100} = 4.5$$

P_{15}th class $= 5 - 10$

$$P_{15} = l + \frac{(15N/100 - m) \times c}{f}$$

$$= 5 + \frac{(4.5 - 4) \times 5}{6}$$

$$= 5 + \tfrac{1}{2} \times 5/6 = 5\tfrac{5}{12}$$

$$P_{22} = 22 \; (30/100) = 6.6$$

This lies in the class 5—10

$$P_{22} = 5 + \frac{(6.6 - 4) \times 5}{6}$$

$$= 5 + \frac{2.6 \times 5}{6} = 5 + 13/6 = 7\tfrac{1}{6}$$

$$P_{70} = \frac{70 \times 30}{100} = 21$$

21st item lies in the class 15—20.

$$P_{70} = 15 + \frac{(21 - 15) \times 5}{7}$$

$$= 15 + \frac{6 \times 5}{7}$$

$$= 15 + 4\tfrac{2}{7}$$

$$= 19\tfrac{2}{7}$$

IV. Calculation from the graph

Percentiles can also be computed from the Ogive curves for the continuous frequency distribution.

Corresponding to the items $N/100$, $2N/100$, $3N/100 \ldots . 7N/100 \ldots . 99N/100$, points on Y-axis can be located. From these points on the Y-axis, we should draw straight lines parallel to the X-axis to cut the Ogive curve at different points say P_1, P_2, P_3, P_4, $\ldots \ldots P_r \ldots \ldots P_{99}$. From these points of intersections on the Ogive curve, we should draw straight lines perpendicular to X-axis or paralllel to Y-axis to cut the X-axis at different points say $X_1, X_2, \ldots \ldots \ldots X_r \ldots \ldots X_{99}$. The values of X corresponding to these points of location on the X axis will give the values of the respective percentiles.

Important points to be noted for calculating Median, Quartiles, Quintiles, Deciles and Percentiles from continuous frequency distribution:

Sometimes, the mid-values or class mark of the each class will be given instead of the classes with their true lower and upper limits. In such cases, the difference between two successive class marks can be taken as the class interval. With the help of these class intervals and the respective mid-values or the class marks, the true lower and upper limits of the classes should be fixed.

From the frequency, we should calculate the cumulative frequency of the distribution.

Properties

1. The values of Medians, Quartiles, Quintiles, Deciles, percentiles should always be expressed in the same unit of measurement as the original units in the distribution.

2. These values are influenced by the location or position of the items and not influenced by the magnitude of values of the items.

LIST OF FORMULAE FOR CALCULATION OF MEASURES OF CENTRAL TENDENCY

Measuresments (1)		Ungrouped data (2)	Discrete frequency distribution (3)	Continuous frequency distribution. (4)
1. Mean		$\Sigma\ x/n$	$\Sigma xf/N$	$\Sigma \frac{x_i f_i}{N}$ (where x_i is the mid-value).
2. Median		$(n+1)/2$th item	$(N+1)/2$th item	$N/2$th item $= l + \frac{(N/2 - m)\,c}{f}$
3. Quartiles	Q_1	$(n+1)/4$th item	$(N+1)/4$th item	$= l + \frac{(N/4 - m)\,c}{f}$
	Q_3	$\frac{3\,(n+1)\text{th}}{4}$ item	$\frac{3\,(N+1)\text{th}}{4}$ item	$= l + \frac{(3N/4 - m)\,c}{f}$
4. Quintiles	1st	$(n+1)/5$	$(N+1)/5$th item	$= l + \frac{(N/5 - m)\,c}{f}$
	2nd	$2(n+1)/5$	$2(N+1)/5$th item	$= l + \frac{(2N/5 - m)\,c}{f}$
	3rd	$3(n+1)/5$	$3\,(N+1)/5$th item	$= l + \frac{(3N/5 - m)\,c}{f}$
	4th	$4(n+1)/5$	$4(N+1)/5$th item	$= l + \frac{(4N/5 - m)\,c}{f}$

(1)		(2)	(3)	(4)
5. Deciles:	D_1	$(n+1)/10$	$(N+1)/10$	$= l + \frac{(N/10-m)\,c}{f}$
		$2(n+1)/10$	$2(N+1)/10$	
D_2	$=$	$(n+1)/5$	$(N+1)/5$	$= l + \frac{(2N/10-m)\,c}{f}$
D_3	$=$	$3(n+1)/10$	$3(N+1)/10$	$= l + \frac{(3N/10-m)c}{f}$
D_4	$=$	$4(n+1)/10$	$4(N+1)/10$	$= l + \frac{(4N/10-m)c}{f}$
		$2(n+1)/5$	$2(N+1)/5$	
D_5	$=$	$\frac{5(n+1)}{10}$	$\frac{5(N+1)}{10}$	$l + \frac{(5N/10-m)\,c}{f}$ (Median)
D_6	$=$	$\frac{6(n+1)}{10}$	$\frac{6\,(N+1)}{10}$	$l + \frac{(6N/10-m)\,c}{f}$
D_7	$=$	$\frac{7\,(n+1)}{10}$	$\frac{7\,(N+1)}{10}$	$l + \frac{(7N/10-m)c}{f}$
D_8	$=$	$\frac{8\,(n+1)}{10}$	$\frac{8(N+1)}{10}$	$l + \frac{(8N/10-m)c}{f}$
	$=$	$\frac{4\,(n+1)}{5}$	$\frac{4\,(N+1)}{5}$	
D_9	$=$	$\frac{9(n+1)}{10}$	$\frac{9\,(N+1)}{10}$	$l + \frac{(9N/10-m)c}{f}$

(1)	(2)	(3)	(4)
6. Percentiles			
P_1	$(n+1)/100$	$(N+1)/100$	$l + \frac{(N/100 - m)\,c}{f}$
P_5	$5(n+1)/100$	$5(N+1)/100$	$l + \frac{(5N/100 - m)\,c}{f}$
P_{10}	$10(n+1)/100$	$10(N+1)/100$	$l + \frac{(10N/10 - m)\,c}{f}$
P_{20}	$20(n+1)/100$	$20(N+1)/100$	$l + \frac{(20N/100 - m)\,c}{f}$
$P_{25} = Q_1$	$25(n+1)/100$	$25(N+1)/100$	$l + \frac{(25N/100 - m)\,c}{f}$
$P_{30} =$	$30(n+1)/100$	$30(N+1)/100$	$l + \frac{(30N/100 - m)\,c}{f}$
P_{40}	$40(n+1)/100$	$40(N+1)/100$	$l + \frac{(40N/100 - m)\,c}{f}$
$P_{50} = Q_2$	$50(n+1)/100$	$50(N+1)/100$	$l + \frac{(50N/100 - m)\,c}{f}$
P_{60}	$60(n+1)/100$	$60(N+1)/100$	$l + \frac{(60N/100 - m)\,c}{f}$
P_{70}	$70(n+1)/100$	$70(N+1)/100$	$l + \frac{(70N/100 - m)\,c}{f}$
$P_{75} = Q_3$	$75(N+1)/100$	$75(N+1)/100$	$l + \frac{(75N/100 - m)\,c}{f}$

(1)	(2)	(3)	(4)
P_{80}	$80(n+1)/100$	$80(N+1)/100$	$l+\dfrac{(80N/100-m)\,c}{f}$
P_{85}	$85(n+1)/100$	$85(N+1)/100$	$l+\dfrac{(85N/100-m)\,c}{f}$
P_{90}	$90(n+1)/100$	$90(N+1)/100$	$l+\dfrac{(90N/100-m)\,c}{f}$
P_{99}	$99(n+1)/100$	$99(N+1)/100$	$l+\dfrac{(99N/100-m)\,c}{f}$

MODE

Another measure of location or central tendency is Mode. Mode is defined as that value which occurs most frequently or typical. Generally Mode represents the value having the largest or the maximum frequency. Mode can be calculated for ungrouped data, for discrete frequency distribution and continuous distribution.

1. Calculation of mode from ungrouped data

Mode can be calculated with ease in the case of ungrouped data. The first step in this process is the re-arrangement of the values in the series in the ascending order of their magnitude. From the series thus arranged, that value which occurs most frequently or that value which occurs the greatest or highest or maximum number of times can be selected as the Mode. Let us see an example:

7, 8, 10, 10, 10, 11, 12, 12, 25, 25, 29.

It is seen certain values for example: 10, 12 and 25 are occuring more than once. While the values 12 and 25 are occuring twice, the value 10 is occuring thrice. Hence the Mode or Modal value for this series is 10.

In the case of Mode, a change in the value of one item can naturally alter the value of Mode. Suppose we replace 10 by 25 in the above series, the series will undergo changes as follows:

7, 8, 10, 10, 11, 12, 12, 25, 25, 25, 29.

Mode in this series—25.

Because of this change in the Modal value due to slight alteration in the value of even one item, it is said that the Modal value is highly unstable or nct steady.

Unimodal

A series may have only one Modal value and it is called Unimodal series.

Example: 7, 8, 10, 10, 10, 11, 12, 12.

Mode = 10.

Bi-modal

Sometimes a series may have two values as modal values and it is said to be a bi-modal set or series.

Example: 7, 8, 10, 10, 10, 11, 12, 12, 12.

In this series, two values namely, 10 and 12 are occuring thrice each. Hence this series is having two Modal values and the Modal values are 10 and 12.

Trimodal

A series which has three Modal values is called tri-modal series. 7, 8, 10, 10, 11, 11, 12, 12.

In this, three values, namely 10, 11 and 12 are occurring each two times. Hence it is a tri-modal series and the Modal values are 10, 11, and 12.

Multi-modal

A series which has more than three Modal values is called a multi-modal series.

2. Calculation of Mode for discrete frequency distribution

In a discrete frequency distribution, the value of the items having the highest or greatest or maximum frequency is taken as the mode.

Example

Value of the items (x)	Frequency (f)
8	4
10	3
12	2
14	7
16	2
18	4
20	1

In this case the value 14 has the greatest frequency namely 7.

Hence Mode = 14.

3. Calculation of Mode for continuous frequency distribution

A. Crude method

1. The maximum frequency of the distribution should be found out first.

2. The class having the highest frequency should then be determined. The class having the highest or maximum frequency is known as Modal class.

3. The mid-value (x_i) of the Modal class should be found out by taking the mid-values of the lower and upper limits of the class intervals. This mid-value is taken as Mode.

The assumption is only an approximation and not actual. Hence the Mode obtained by this method is only an approximate value and not an accurate value.

Example (1)

Weight in kg. (x)	Frequency (f)
40 – 50	4
50 – 60	6
60 – 70	7
70 – 80	12
80 – 90	4
90 – 100	6
100 – 110	8

Maximum frequency = 12.

Modal class = 70 – 80.

Mid-value of the modal class $= \dfrac{70 + 80}{2}$

$= 150/2 = 75$ kg.

$\therefore$ Mode = **75 kg.**

Example (2)

Life in hours (x)	No. of tube lights (f)
400 — 700	7
700 — 1000	10
1000 — 1300	4
1300 — 1600	14
1600 — 1900	6
1900 — 2200	4
2200 — 2500	2

Highest frequency $= 14.$

Modal class $= 1300 - 1600$

Mid-value $= \dfrac{1300 + 1600}{2} = \dfrac{2900}{2} = 1450.$

Mode = **1450** hours

B. Calculation of Mode by giving weights to the preceeding class and succeeding class of the modal class

The value of the mode is sometimes affected or influenced by the frequencies in the preceeding and succeeding classes of the modal class.

If the frequency of the preceeding class is greater than the frequency of the succeeding class, the value of the Modal class will be nearer to the lower limit of the Modal class instead of concentrating on the mid-value of the Modal class.

On the other hand, if the frequency of the succeeding class is greater than the frequency of the preceeding class then the value of the Mode will be nearer to the upper limit of the Modal class instead of concentrating on the Mid-value of the modal class.

Therefore, in calculating the value of the Mode, the frequencies of the preceeding and succeeding classes are also taken into consideration.

Example (1)

Weight in kg. (x)	No. of bundles (f)
40 — 50	4
50 — 60	6
60 — 70	7
70 — 80	12 ——— X
80 — 90	4
90 — 100	6
100 — 110	8

The modal class is 70-80 since it is having the greatest frequency namely 12.

Lower limit of the modal class $= l = 70$.

Preceeding class 60 — 70.

Frequency of the preceeding class = 7.

Let us denote it by the letter f_1. $\therefore$ $f_1 = 7$.

Succeeding class 80—90.

Frequency of the succeeding class = 4.

Let it be denoted by the letter f_2. $\therefore$ $f_2 = 4$.

Width of the Modal class $= 80 - 70 = 10$.

The following formula is adopted for calculation of Mode.

$$\text{Mode} = l + \frac{cf_2}{f_1 + f_2}$$

Let us substitute the value in the formula.

$$\text{Mode} = 70 + \frac{10 \times 4}{7 + 4}$$

$$= 70 + \frac{40}{11} = 70 + 3\frac{7}{11} = 73\frac{7}{11} \text{ or } 73.64 \text{ kg.}$$

Example (2)

Life in hours (x)	No. of tube lights (f)
400 — 700	7
700 — 1000	10
1000 — 1300	4
1300 — 1600	14
1600 — 1900	6
1900 — 2200	4
2200 — 2500	2

Maximum frequency = 14.

Modal class = 1300 — 1600.

l — true lower limit of the modal class = 1300.

c — width of the modal class = 1600 — 1300

= 300.

f_1 — frequency of the preceeding class $= 4$.

f_2 — frequency of the succeding class $= 6$.

$$\text{Mode} = l + \frac{c.\, f_2}{f_1 + f_2} = 1300 + \frac{300 \times 6}{4 + 6}$$

$$= 1300+1800/10$$

$$= 1300+180$$

$$= 1480 \text{ Hours.}$$

C. Calculation of Mode by taking the differences in the frequencies between the Modal class and the preceeding class: 2) Modal class and the succeeding class

The above formula undergoes a slight change as follows:

$$\text{Mode} = l + \frac{c.\, d_1.}{d_1 + d_2}$$

l = True lower limit of the Modal class.

c = Width or class interval of the Moda lclass.

d_1 = difference between the frequencies of Modal class and the preceeding class. (Only absolute value without sign is considered.) $14 - 4 = 10$.

d_2 = difference between the frequencies of modal class and succeeding class. $14 - 6 = 8$.

Substituting these values in the formula we get,

$$\text{Mode} = l + \frac{c.\, d_1}{d_1 + d_2}$$

$$\text{Mode} = 1300 + \frac{300 \times 10}{10 + 8}$$

$$= 1300+3000/18$$

$$= 1300+166.67$$

$$= 1466.67 \text{ Hours.}$$

Note: Of these two formulae the second formula gives more accurate result and hence it is preferable:

1. Mode : $l + \dfrac{c \times f_2}{f_1 \times f_2}$

2. Mode : $l + \dfrac{c.\ d_1}{d_1 + d_2}$

A slight change in the second term of the formula may be noted:

(1) $\dfrac{c.\ f_2}{f_1 + f_2}$ (2) $\dfrac{c.\ d_1}{d_1 + d_2}$

In the first case, the frequency of the modal class is not at all considered while in the second case the frequency of the mcdal class is also considered indirectly by calculating the difference of the frequencies of the neighbouring classes from the frequency of the Modal class. Hence the second formula gives better results.

(1) Actual frequencies in the preceeding and succeeding classes are used in the first formula. In the second formula, the differences in the frequencies compared with the frequency of the Modal class are used.

(2) In the numerator of the first formula the actual frequency of the succeeding class is used. But in the case of the second formula, the difference of the frequency of the **preceeding** class is used.

3. One practical difficulty may arise in using these two formulae. Sometimes the Modal class may happen to be either first or last class. In such rare situations we cannot have either preceeding or succeeding class. In all such rare cases, it is better to take the mid-value of the Modal class itself as the Mode.

D. Calculation of Mode from Mean and Median

Mode can also be calculated from the two other measures of central tendency namely, Mean and Median. In symmetric

distributions, the frequencies on either side of the Mean will be more or less identical. An example of one such distribution is given below:

(x)	(f)
0 — 5	2
5 — 10	3
10 — 15	4
15 — 20	10
20 — 25	4
25 — 30	3
30 — 35	2

In the above distribution, the class interval is uniform. The maximum frequency is 10 in the class 15—20. On either side of this class, the frequencies are identical. This is called a *symmetrical distribution.*

In such symmetrical distributions, Mean and Median will have the same value. In other words, Mean and Median will try to approach each other and consequently their difference will be very small. Hence the difference between Mean and Median can be taken as a measure for ascertaining the symmetry. But the situation is different in the case of Mode. The difference between Mean and Mode will be greater than that of Mean and Median. The difference between Mean and Mode (Mean — Mode) is compared with the difference between Mean and Median (Mean — Median). It is computed that the difference between Mean and Mode (Mean — Mode) is equal to **thrice** the difference between Mean and Median, that is,

3 (Mean—Median). Hence this is used as a formula to determine the value of mode.

$$\text{Mean} - \text{Mode} = 3\ (\text{Mean} - \text{Median})$$

$$\therefore\ \text{Mode} = 3\ \text{Median} - 2\ \text{Mean}.$$

Example: Calculate the Mode

(1) Mean $= 125$: Median $= 115$.

$$(\text{Mean} - \text{Mode}) = 3\ (\text{Mean} - \text{Median})$$

$$(125 - \text{Mode}) = 3\ (125 - 115)$$

$$= 3 \times 10 = 30.$$

$$\therefore\ \text{Mode} = 95$$

(2) 1 Mode $= 3$ Median $- 2$ Mean.

$$= 3 \times 115 - 2 \times 125$$

$$= 345 - 250$$

$$= 95.$$

E. Determination of Mode from graph

Mode can also be determined from graph. In this connection, the students should note one basic difference. While the other measures like Median, Quartiles, Quintiles, Deciles and Percentiles are calculated from the cumulative frequencies, the Mode is calculated from the actual frequency only. Since cumulative frequencies are used for the calculation of other measures namely, Median, Quartiles, Quintiles, Deciles and Percentiles the Ogive curve for less than cumulative frequencies is used for calculation. As Mode has to be calculated from the actual frequency, the curve for the frequency, namely, the frequency curve has to be used for Mode.

In a frequency curve, the X-axis will represent the value and the Y-axis will represent the frequency. In other words, X-coordinate will represent the value of items and the Y co-ordinate will represent the frequency. We should draw different 'Y'

co-ordinates for different X co-ordinates. That value of X co-ordinate for which the 'Y' co-ordinate is the maximuum is taken as the Mode.

Through the apex (highest point) of the frequency curve, a vertical line perpendicular to the X-axis to cut the X-axis should be drawn. The point of intersection of the vertical line with the X-axis may be noted. The value of X, corresponding to this point of intersection will represent the Mode.

Ungrouped data	Discrete frequency distribution	Continuous frequency distribution
1. Values have to be rearranged in the ascending order.	Values of the items have to be arranged in ascending order.	1. Mid value of the modal class.
2 Value of the items in the array which occurs for greatest number of times.	Value of the items having the highest frequency.	1. $l + \frac{c \cdot f_2}{f_1 + f_2}$ 2. $l + \frac{c \cdot d_1}{d_1 + d_2}$ 3. Mean— Mode = 3 (Mean-Median) 4. 1Mode = 3Median 2 Mean.

GEOMETRIC MEAN (G.M.)

We have already studied the Arithmetic Mean. In this section we shall study another mean called Geometric Mean.

While the arithmetic mean is calculated from the total sum of the values of the items, the Geometric Mean is calculated from **the product of the values of the items.**

While the arithmetic mean is calculated from the sum by dividing it by the number of items, the Geometric Mean is calculated from the product by finding the root corresponding to the number of items.

A. Calculation of Geometric Mean for ungrouped data

1. Let us calculate the Geometric Mean for the following items 9 and 16.

Number of items = 2.

Product of the items = 9 × 16 = 144.

Since there are only two items, we should find the square root of the product $2\sqrt{144} = 12$.

∴ Geometric Mean = 12.

2. Calculate the Geometric Mean for the following items;

4, 16 and 8.

Number of items = 3.

Product ofthe items = 4 × 16 × 8 = 512.

∴ Geometric Mean = $3\sqrt{512} = 8$.

General Formula

The general formula can be derived:

Suppose there are '*n*' items, say

$x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n.$

$$\text{G.M.} = \sqrt[n]{x_1 \times x_2 \times x_3 \times x_4 \ldots\ldots\ldots \lambda_n}$$

$$\text{(or) G} = \sqrt[n]{x_1 \times x_2 \times x_3 \times x_4 \times_5 \ldots\ldots\ldots x_n}$$

Shortcut Method

If there are more items, finding the product by multiplication is more tedious. Finding the *n*th root of the product is still more

laborious. Hence we have to find out shortcut method. The shortcut method for multiplication is use of logarithm. Therefcre, we can use logarithms for finding the Geometric Mean.

Those who are not familiar with logarithms may follow the illustration given below:

$$a^3 \times a^1 = a^{3+4} \text{ or } a^7 \text{ and not } a^{3\times4} \text{ or } a^{12}.$$

This is because a^3 and a^4 can be written as follows:

$$a^3 = a \times a \times a \quad \text{and } a^4 = a \times a \times a \times a.$$

$$\therefore \quad a^3 \times a^4 = (a \times a \times a) \times (a \times a \times a \times a) = a^{3+4} = a^7$$

Here the figures are expressed in terms of powers of 'a'.

Let us substitute the value 10 for 'a'.

$$a^3 = 10^3 = 10 \times 10 \times 10 = 1000.$$

$$a^4 = 10^4 = 10 \times 10 \times 10 \times 10 = 10{,}000$$

So, $1000 \times 10{,}000$ can be written as $10^3 \times 10^4 = 10^{3+4} = 10^7$.

The powers are called logarithms.

$$\therefore \quad Log_{10}\,(1000 \times 10{,}000) = log_{10}\,1000 + log_{10}\,10{,}000$$
$$= 3 + 4 = 7$$

From this, it can be seen that the powers of the product of any two numbers will be equal to the sum of the powers of the numbers when the power is expressed in terms of a common value as base. The same principle applies to the product of any number of Numbers. From the power of the product we can find out the product itself. The power is called the logarithm and the product is called anti-logarithm. The common term in which the power is expressed is called the base.

In order to find out the product of different numbers, we use the following procedure:

1. We first find out the logarithm of each number.
2. After finding out the logarithms, we should add and find out the sum of the logarithms.
3. We find the anti-logarithm for the sum of the logarithms and the antilogarithm would represent the product.

The same procedure can be followed to find out the Geometric Mean.

1. Find out the logarithm of each value.
2. Add the logarithms and find the sum.
3. Divide the sum of the logarithms by the number of items in the series. (Find the Arithmetic Mean of the logarithms. This Arithmetic Mean of the logarithms would represent the logarithm of Geometric Mean.)
4. Lastly, find out the anti-logarithm for the Arithmetic Mean of the logarithms which will be equal to the G.M.

Formula

The formula can be derived as follows:

Let there be 'n' values of x and they may be represented by $X_1, X_2, X_3 \ldots\ldots\ldots\ldots X_n$.

Sum of the logarithms $= log\ x_1 + log\ x_2 + log\ x_3 + \ldots + log\ x_n$

$$x_n\ (or) = \Sigma\ log\ x$$

Arithmetic Mean of the logarithms $= \dfrac{log\ x_1 + log\ x_2 + log\ x_3 + \ldots log\ x_n}{n}$

$$= \frac{\Sigma\ log\ x}{n}$$

$$\therefore\ log\ (\text{G.M.}) = \frac{\Sigma\ log\ x}{n}$$

$$\therefore\ \text{Geometric Mean} = \text{Antilog of } \frac{\Sigma\ log\ x}{n}$$

Let us calculate the Geometric Mean of the following items: 1225, 148, 79, 1478 and 9.

$$\begin{aligned}
\log 1225 &= 3.0881 \\
\log 148 &= 2.1703 \\
\log 79 &= 1.8976 \\
\log 1478 &= 3.1697 \\
\log 9 &= 0.9542 \\
\text{Total} &= \mathbf{11.2799}
\end{aligned}$$

$$n = 5. \quad \text{Mean} = \frac{11.2799}{5}$$

$$= 2.2559 \text{ or } 2.2560.$$

Antilog of 2.2560 $=$ **180.3**

$\therefore$ Geometric Mean $=$ **180.3**

B. Calculation of Geometric Mean from Discrete frequency distribution

Let us calculate the Geometric Mean of the following values:

Value		Frequency	
3	x_1	2	f_1
4	x_2	3	f_2
6	x_3	1	f_3
8	x_4	4	f_4
	Total	10	

This can be re-written as follows:

3, 3, 4 4, 4, 6, 8, 8, 8, 8.

(1)		(2)
log 3		= 0.4771
log 3		= 0.4771
log 4		= 0.6021
log 4		= 0.6021
log 4		= 0.6021
log 6		= 0.7782
log 8		= 0.9031
log 8		= 0.9031
log 8		= 0.9031
log 8		= 0.9031
	Total	7.1511

We can adopt still shorter method in totalling the logarithms.

The value 3 is repeated two times. Therefore instead of writing the *log* (3), twice, we can multiply the *log* (3) by 2. In the same manner we can multiply *log* (4) by 3, *log* (6) by 1 and *log* (8) by 4 and calculate the total of logarithms as follows:

$2 \times \log 3 = 2 \times 0.4771 = 0.9542$

$3 \times \log 4 = 3 \times 0.6021 = 1.8063$

$1 \times \log 6 = 1 \times 0.7782 = 0.7782$

$4 \times \log 8 = 4 \times 0.9031 = 3.6124$

7.1511

After finding the total, we can find the average of the logarithms.

Since there are 10 items, we can divide the total by 10.

$$\therefore \quad \frac{7.1511}{10} = 0.7151$$

$$Log\ (\text{G.M.}) = 0.7151$$

$$\text{G.M.} = \text{antilog of } 0.7151$$

$$= 5.189.$$

$\therefore$ The formula can be written as follows:

$$\frac{2 \times log\ 3 + 3 \times log\ 4 + 1 \times log\ 6 + 4 \times log\ 8.}{2 + 3 + 1 + 4}$$

If we replace the values by x_1, x_2, x_3 and x_4 and the frequencies are replaced by f_1, f_2, f_3 and f_4 the formula would emerge as follows:

$$Log\ (G.M.) = \frac{f_1 \times log\ x_1 + f_2\ log\ x_2 + f_3\ log\ x_3 + f_4\ log\ x_4}{f_1 + f_2 + f_3 + f_4}$$

$\therefore$ If there are 'n' items the general formula would be,

$$log\ G = \frac{f_1\ log\ x_1 + f_2\ log\ x_2 + \ldots\ldots + f_n\ log\ x_n}{f_1 + f_2 + f_3 + \ldots\ldots\ldots + f_n}$$

$$= \frac{\Sigma\ f_i\ log\ xi}{\Sigma\ fi}$$

$$\therefore \quad \text{Geometric Mean} = \text{Anti log } \frac{\Sigma\ f\ log\ x}{n}$$

Therefore the previous example can be worked out as follows:

Value (x)	Frequency (f)	Log x	f *Log* x
3	2	0.4771	0.9542
4	3	0.6021	1.8063
6	1	0.7782	0.7782
8	4	0.9031	3.6124
Total	10		7.1511

$$\text{Average of log} = \frac{7{\cdot}1511}{10}$$

$$= 0{\cdot}7151$$

$$\text{Log. of Geometric Mean} = 0{\cdot}7151$$

$$\therefore \text{ G.M.} = \text{anti log } (0{\cdot}7151)$$

$$= 5{\cdot}189$$

C. Calculation of Geometric Mean from continuous frequency distribution

We shall consider the following frequency distribution.

Class interval	Frequency
0 — 10	4
10 — 20	8
20 — 30	10
30 — 40	5
40 — 50	3
Total	30

1. We should first find out the mid-value of each class. Afterwards the proc dure will be similar to the method adopted in the case of discrete frequency distribution. The distribution can be written as follows:

Mid values (x_i) (1)	Frequency (f) (2)	log (x) (3)	$f \times log\ (x)$ (4)
5	4	0.6990	2.7960
15	8	1.1761	9.4088
25	10	1.3979	13.9790
35	5	1.5441	7.7205
45	3	1.6532	4.9596
Total	30		38.8639

$$N = \Sigma f = 30.$$

$$\Sigma f.\ log\ x = 38.8639$$

$$log\ (G.M.) = \frac{\Sigma f.log\ x}{N} = \frac{38.8639}{30} = 1.2955$$

$$G.M. = \text{Antilog}\ (1{\cdot}2955)$$

$$= 19.74$$

Uses of Geometric Mean

1. Geometric mean is a better average to indicate the rate of change. When percentage increases over a period of time are given, we must use only Geometric Mean to find out the average precentage increase.

2. Geometric mean is mostly used in cases where the variation in the values over a period of time takes place at a compound rate as in the case of investment at compound interest or in the growth of population.

Example

(1) Where percentage increases are given:

The population of a country has increased by 15% from 1941 to 1951. It has increased by 20% from 1951 to 1961, by 25% from 1961 to 1971. Calculate the average increase per decade in the population.

We should assume the population at the beginning of each decade as 100 and proceed further:

1. Population at the beginning of the first decade (1941–1951)	=	100.
Population at the end of the first decade	=	115.
	=	100 + 15
2. Population at the beginning of the second decade (1951–61)	=	100
Population at the end of the second decade	=	120
	=	100 + 20
3. Population at the beginning of the third decade (1961–71)	=	100
Population at the end of the third decade	=	125
	=	100 + 25

Geometric Mean of the population at the end of each decade is obtained by using the following formula:

$$\text{Geometric Mean} = \text{Anti log} \frac{\Sigma \log x}{n}$$

$$n = 3.$$

$$\log x_1 = \log 115 = 2.0607$$
$$\log x_2 = \log 120 = 2.0792$$
$$\log x_3 = \log 125 = 2.0969$$

Total $= 6.2368$

$$\frac{\Sigma \log x}{n} = \frac{6.2368}{3}$$

$$= \mathbf{2.0789}$$

Geometric Mean $=$ Anti log (2.0789)

$= 119.9$

Rate of increase $= 119.9 - 100.0$

$= \mathbf{19.9}$

Absolute figure

Sometimes absolute figures will be given. In such cases we should first convert the absolute figures into percentages and proceed afterwards as before.

Year	Population in millions
1901	200
1911	225
1921	260
1931	290

Here absolute figures are given:

1. If the population of 1901 is taken as 100, the population of 1911 will be

$$\frac{225}{200} \times 100 = 112{\cdot}5$$

2. If the population of 1911 is taken as 100, the population of 1921 will be

$$\frac{260}{225} \times 100 = \frac{1040}{9} = 115{\cdot}6$$

3. If the popuation of 1921 is 100, the population of 1931 will be:

$$\frac{290}{260} \times 100 = 111{\cdot}5$$

The relative values are,

112·5, 115·6 and 111·5

n = Number of decades = 3.

$$\begin{array}{lll} 112{\cdot}5 & \log(112{\cdot}5) & = 2{\cdot}0511 \\ 115{\cdot}6 & \log(115{\cdot}6) & = 2{\cdot}0630 \\ 111{\cdot}5 & \log(111{\cdot}5) & = \underline{2{\cdot}0472} \\ & & \ 6{\cdot}1613 \end{array}$$

$$\text{Log (G.M.)} = \frac{6{\cdot}1613}{3}$$

$$= 2{\cdot}0538$$

$$\text{Anti log}\,(2{\cdot}0538) = 113{\cdot}2$$

$$\therefore \text{ Increase in population} = 113{\cdot}2 - 100{\cdot}0$$

$$= \mathbf{13{\cdot}2\%}$$

We shall use Geometric Mean in cases where the variations in the values take place at a compound rate as in the case of compound interest or in the growth of population. In such cases, the following formula can be adopted. This is nothing new, since the students would have studied it in the lower class when they studied compound interest. The only difference is that we have used logarithms at present.

$$P_n = P_0\,(1 + r)^n$$

Where P_n = Value at the end of the nth period.

P_o = Value at the beginning of the period.

r = rate of change

n = the number of years.

If we know P_o and P_n, and 'n' (since we are interested in the average) we can calculate the average of P_n and P_o.

Example

The population of Tamil Nadu in 1961 and 1971 are given below and calculate the average population:

1961 = 337 millions

1971 = 411 millions

n = 2 (number of items)

log 337 = 2.5276

log 411 = 2.6138

5.1414

$$\text{Log}\,(GM) = \frac{5.1414}{2} = 2.5707$$

$\therefore$ GM = anti log (2 5707)

= **372.1**

Instead of population, we can use the same figures as Rupees in lakhs.

1961 ———— 337 Rs. in lakhs.

1971 ———— 411 Rs. in lakhs.

Even then, the methods are the same and the result is also the same.

Formula used

1. Ungrouped data $= G =$ Anti log $\dfrac{\Sigma \log x}{n}$

2. Discrete frequency distribution $= G =$ Antilog $\dfrac{\Sigma f . \log x}{N}$

3. Continuous frequency distribution $=$ Anti log $\dfrac{(f . \log xi)}{N}$

Where xi stands for the mid-values of the class interval.

Difference between Arithmetic Mean and Geometric Mean

Arithmetic Mean	Geometric Mean
1. Sum of the values divided by the number of items say 'n' gives the Arithmetic Mean.	The nth root of the product of the values is the Geometric Mean.
2. The average multiplied by 'n' or 'n' times the average will give the total value of all the items $\Sigma x_i = n \times \bar{x}$.	The Geometric Mean raised to the power 'n' $(GM)^n$ gives product of all the values $(GM)^n = \pi\, (x_i)$ where π stands for the symbol of multiplication.
3. All distributions having the equal number of items and also having the same total value will have the same average even though individual value of one distribution does not agree with the counterpart of the other.	All series having equal number of items and the same product value will have the same Geometric Mean even though individual value of one distribution does not agree with the counterpart of the other distribution.
4. Even if the value of one of the items is 0, the Arithmetic Mean can be **calculated.**	If the value of one of the items is 0, the product of all the values will be '0' and consequently, the **Geometric Mean will be 0.**

Arithmetic Mean	Geometric Mean
5. Even if the value of one of the items is negative, the Arithmetic Mean can be calculated.	If the value of one of the items is negative (-ve), the product of all the items will be negative and consequently the Geometric Mean is imaginary.

HARMONIC MEAN (H.M)

Though the Arithmetic Mean is based on the given values the Harmonic Mean is calculated on the basis of the reciprocals of the given values.

Calculation of harmonic mean for ungrouped values

Let us consider the values: 4, 5 6.

(1) We must find out the reciprocals of each of the values:

Value	Reciprocal
4	$\frac{1}{4}$
5	1/5
6	1/6

Note: The product of a given value and its reciprocal will always be 1. $\frac{1}{4}$ is the reciprocal of 4 and 4 is the reciprocal of $\frac{1}{4}$, since their products in both the occassions are 1.

(2) We must find out the total value of the reciprocals.

$$\frac{1}{4} + \frac{1}{5} + \frac{1}{6} = \frac{30 + 24 + 20}{120} = \frac{74}{120} = \frac{37}{60}$$

(3) We must calculate the arithmetic mean of the sum of their reciprocals by dividing the total value of the reciprocals by the number of items.

$$\text{Mean of the reciprocals} = \frac{\text{Total of reciprocals}}{\text{No. of items}}$$

$$= \frac{37}{60} \times \frac{1}{3}$$

$$= \frac{37}{60 \times 3} = 37/180.$$

(4) Find out the reciprocal of the Mean of the reciprocals which is equal to Harmonic Mean.

Mean of the reciprocals $= 37/180$.

$$\left.\begin{array}{r}\text{Reciprocal of the Mean of the}\\ \text{reciprocals}\end{array}\right\} = \frac{1}{\frac{37}{180}} = \frac{180}{37}$$

$$\therefore \quad \text{Harmonic Mean} = 4\,\frac{32}{37}$$

Definition of Harmonic Mean

Now we can define Harmonic Mean. It is the reciprocal of the Arithmetic Mean of the reciprocals of the given values.

Let us examine the formula with an example.

Find out the Harmonic Mean of the following values:

20 25 30 35, 40.

Their reciprocals are:

1/20,1/25, 1/30, 1/35, 1/40.

Values	Reciprocals
$x_1 = 20$	$1/x_1 = 1/20$
$x_2 = 25$	$1/x_2 = 1/25$
$x_3 = 30$	$1/x_3 = 1/30$
$x_4 = 35$	$1/x_4 = 1/35$
$x_5 = 40$	$1/x_5 = 1/40$

Number of items $= N = 5.$

Sum of the reciprocals $= \dfrac{1}{x_1} + \dfrac{1}{x_2} + \dfrac{1}{x_3} + \dfrac{1}{x_4} + \dfrac{1}{x_5}$

$= 1/20 + 1/25 + 1/30 + 1/35 + 1/40.$

Arithmetic Mean of the reciprocals $= \dfrac{1/20 + 1/25 + 1/30 + 1/35 + 1/40}{5}$

Reciprocal of the A.M. $= \dfrac{5}{1/20 + 1/25 + 1/30 + 1/35 + 1/40}$

$= \dfrac{n}{1/x_1 + 1/x_2 + 1/x_3 + 1/x_4 + 1/x_5}$

Harmonic Mean $= n \,/\, \Sigma\, 1/x$

$= \dfrac{5}{0.050 + 0.040 + 0.033 + 0.029 + 0.025}$

$= 5/0.177 =$ **28.25**

Harmonic Mean $= n \,/\, \Sigma\, 1/x$

Calculation of Harmonic Mean from discrete frequency distribution

In the previous cases, each value occurs only once. In other words, the frequency of the value is one in each case. So the reciprocal in the previous example can be interpreted in the light of the frequency ie: the reciprocal is the result obtained by dividing the frequency by the given value $\left(\frac{f}{x}\right)$. This principle or interpretation is applied in the case of discretefrequency distribution. Afterwards, the formula would undergo changes as follows:

Original formula for ungrouped data $= n \,/\, \Sigma f/x$

Formula for discrete frequency distribution $= \dfrac{\Sigma f}{\Sigma f/x} = \dfrac{n}{\Sigma f/x}$

Each frequency has to be divided by the respective value and this would be same as the reciprocal of the values.

Example

Value (x)		Frequency (f)	Reciprocals f/x	
10		10	10/10	= 1.00
20		15	15/20	= 0.75
30		40	40/30	= 1.33
40		10	10/40	= 0.25
50		5	5/50	= 0.10
	Total	80		3.43

Mean of the reciprocal = 3.43/80

Reciprocal of the mean = 80/3.43

Harmonic Mean (or) H = 23.3

Calculation of Harmonic Mean from continuous frequency distribution

While in the case of discrete frequency distribution, the actual values are given, the classes with their lower and upper li rits will be given in the case of continuous frequency distribution.

Thereforе, we should first calculate the mid value of each of the classes. Afterwards, the procedure will be as before.

Class	Frequency
0 – 10	5
10 – 20	7
20 – 30	12
30 — 40	4
40 — 50	2
	30

The above table will be replaced by the following table:

Mid value x		Frequency f	f/x		
5		5	5/5	=	1.00
15		7	7/15	=	0.47
25		12	12/25	–	0.48
35		4	4/35	=	0.11
45		2	2/45	=	0.04
	Total	30			2.10

Mean = 2.10/30 Formula: $\dfrac{\Sigma fi}{\Sigma \dfrac{fi}{xi}} = \dfrac{N}{\Sigma \dfrac{fi}{xi}}$

Reciprocal = 30/2.10 Where xi stands for the mid value of the class

= 14.29

Uses of Harmonic Mean

In this context it must be clearly understood that the original value cf any nilirn and its reciprocal are inversely proportional to each other which neans that their prcduct will always be equal tc 1. This indirectly indicates that Harmonic Mean can be applied in cases of changes taking place in inverse proportion. This can further be explained with the help of an example.

Let us consider the question of number of workers and their wages. These two variables, namely the number of workers and wages or the total number of workers and their total wages are directly proportional. In other words, if the value of one increases the value of other item will also consequently increase or if the value of one decreases the value of other will also deciease.

Let us consider a problem which is very common. The distance between two places namely A and B is 24 km. A cyclist is travelling from A to B at an average speed of 4 km. per hour and returns from B to A at an average speed of 6 km. per hour. What is his average speed?

Anyone will be tempted at the first sight to say that the average speed is 5 km. per hour.

ie : $\dfrac{4+6}{2}$ = 10/2 = 5 km. which is not correct.

An exercise of slight imagination will explain this.

The distance between A & B = 24 km.

The average speed of the journey from A to B = 4 km. per hour.

∴ Time taken to travel from A to B = 24/4 = 6 hrs.

The average speed of the journey from B to A = 6 km. per hour.

∴ Time taken to travel from B to A = 24/6 = 4 hrs.

Total distance travelled = 24 + 24 = 48 km.

Total time taken = 6 + 4 = 10 hrs.

$$\therefore \text{ Average speed } = \frac{\text{Total distance travelled}}{\text{Total time taken.}}$$

$$= 48/10 = 4.8 \text{ km. per hour.}$$

The difference between the first answer and the subsequent answer may be noted. The correctness of the second answer needs no special emphasis. The difference is due to the fact that the two variables or factors namely speed and time are inversely proportional. As the speed increases the time taken will decrease or as the speed decreases the time taken will increase.

The scope for the application of the Geometric Mean or Harmonic Mean is very limited. We mostly use only Arithmetic Mean. Therefore, we shall concentrate our study more on the Arithmetic Mean and its computation.

List of formulae for Harmonic Mean

1. For ungrouped data $= \quad H = \dfrac{n}{\Sigma\, 1/x}$

2. For discrete frequency distribution $H = \dfrac{\Sigma f}{\Sigma f/x}$

$$= \frac{N}{\Sigma\, f/x}$$

3. For continuous frequency distribution $= \dfrac{\Sigma\, f_i}{\Sigma\, fi/xi}$

$$= \frac{N}{\Sigma\, f/x_i}$$

Where x_i stands for the mid-value.

Relationship between Arithmetic Mean, Geometric Mean and Harmonic Mean

For a given set of values, (1) Arithmetic Mean will be greater than or equal to Geometric Mean; (2) Geometric Mean will be greater than or equal to Harmonic Mean.

In other words, for a given set of values the Arithmetic Mean is greater than or equal to Geometric Mean which in turn is greater than or equal to Harmonic Mean.

Arithmetic Mean $\geq$ Geometric Mean $\geq$ Harmonic Mean.

The students would have already studied this property when they studied Progressions in Algebra.

Weighted Mean

When we study about the different means, namely Arithmetic Mean, Geometric Mean, and Harmonic Mean, we have considered three types under each case.

1. Ungrouped data
2. Discrete frequency distribution
3. Continuous frequency distribution

It can be broadly classified under two kinds.

1. Ungrouped data (which involves no frequency)
2. Frequency distribution.

In the case of ungrouped data, each value is considered only once. In other words, the frequency of each value is the same or uniform. When we say the frequencies of all the values are same, it does not mean that the frequency in each case is 1. But the frequency may be 1 or any other value. But whatever may be the value of frequency, it is uniformly same for all the values.

Simple Mean: Whenever the Mean (Arithmetic Mean or Geometric Mean or Harmonic Mean) is calculated without reference to the frequency it is called Simple Mean.

Weighted Mean: Whenever the Mean (Arithmetic Mean or Geometric Mean or Harmonic Mean) is calculated with reference to the frequency it is called weighted mean. In the case of Weighted Mean, the frequency is serving as the weight. In other cases, as in the case of average yield of produce in a district, the area under the crop in different taluks may serve as the weight.

However, the simple mean and the weighted mean are one and the same, when the frequencies in all cases are same or uniform. Let us examine this:

Value (x) (1)	Frequency (f) (2)	$f.x.$ (3)
10	5	50
30	5	150
40	5	200
50	5	250
130	20	650

$$\text{Mean} = \frac{\Sigma\, x.f}{\Sigma f} = \frac{\Sigma\, fx}{N} = \frac{650}{20} = 32.5$$

If we add all the values given in the col (1), we get 130.

The No. of items = 4.

$$\therefore \text{ Average} = \frac{130}{4} = \mathbf{32.5}$$

We find that the means calculated in both the ways ie: with frequency and without frequency, are one and the same.

$$\text{Weighted Mean} = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

$$= \frac{(10\times 5) + (30\times 5) + (40\times 5) + (50\times 5)}{(5) + (5) + (5) + (5)}$$

This can be simplified by taking 5 outside the bracket in the numerator and in the denominator.

$$= \frac{5\,(10 + 30 + 40 + 50)}{5\,(1 + 1 + 1 + 1)}$$

Cancelling 5 both in the Numerator and Denominator we get,

$$\frac{10 + 30 + 40 + 50}{1 + 1 + 1 + 1} = 130/4 = \frac{\Sigma x}{n} = \mathbf{32.5}$$

Therefore, what we have studied under discrete and continuous frequency distribution under all categories of means are nothing but weighted mean, whereas what we have studied under ungrouped data is simple mean.

Comparative merits of different measures of Central tendency

All Means

They have well defined formulae. They are easily amenable for algebraic treatment. while the Arithmetic Mean can be easily computed, the computation of Geometric Mean and Harmonic Mean will involve certain difficulties. While the Arithmetic Mean can be easily understood, the other two require some imagination. These measures cannot be compiled from the graph. All the measures are expressed in the same unit as the original units. They are easily influenced by extreme low or extreme high values. They are not affected by position of the values. Therefore, re-arrangement cf the data either in the ascending order or descending order is not necessary for computation. Cumulative frequencies cannot be used.

Median, Quartiles, Quintiles, Deciles and Percentiles

These have well defined formulae. They have relation to each other.

$$M = Q_2 = D_5 = P_{50}.$$

$$Q_1 = P_{25}$$

$$Q_3 = P_{75}$$

They can be easily understood. They can also be computed from the graph or the cumulative frequency distribution. Cumulative frequencies are used. They are not influenced by the values. Therefcre, extreme high or low values have no influence on them. They are influenced by the position. Therefore, arrangement of data in the order of magnitude is required. They are also expressed in the same units as the original values. They have no direct relationship to Means. Of course in the case of symmetrical distribution, Median will be equal to Arithmetic Mean.

Mode

It has its own formulae for computation. But both of them are crude formulae. This is also not affected by the values but only influenced by the position of the item. Extreme values have no influence. It can be computed from the frequency distribution. It is also expressed in the same unit as the original unit. Introduction of new items may alter the position. There may be more than one Mode for a same distribution. The distribution may be unimodal or bi-modal or trimodal or multi-modal. Arrangement of data in the order of magnitude is necessary. It can also be computed from graph. It has some direct relationship with the Arithmetic Mean and Median. In other words, it can be computed from Arithmetic Mean and Median by using any of the two formulae:

(1) Mean — Mode = 3 (Mean — Median)

(2) 1 Mode = 3 Median — 2 Mean.

Exercises

(1) Calculate the Mean, Median, Mode and Quartiles for the following data.

	fi
0 – 10	3
10 — 20	5
21 — 30	7
31 — 40	8
41 — 50	2
	25

(2) Calculate the average for the combination.

	No. of students	Average weight
Group A	28	45 kg.
Group B	42	43 kg.

(3) Calculate the Mean and Median.

x	f
10	2
15	21
20	25
25	17
30	5

(4) Calculate the different measures of Central tendencies.

10, 12, 15, 29, 40, 12, 13, 15, 16, 17.

(5) Calculate the G.M. of the following data.

(i)	(ii)	(iii)
10	120	1200
15	280	2800
18	320	3200
20	420	4200
25	550	5500

(6) Calculate the G.M. for the following data.

(i) x	f	(ii) x	f
10	4	120	4
15	3	150	3
18	5	180	5
20	6	290	3
25	2	350	5

(7) Calculate the H.M. for the following data.

10	120
15	150
20	180
25	200
30	250

(8) Calculate the H.M. for the following data.

(*i*) x	f	(*ii*) x	f
10	3	150	3
15	4	180	4
18	3	200	2
20	2	225	3
25	1	250	3

(9) Calculate the P_7 P_{25} for the following data.

x	f
0 – 10	5
10 – 20	8
20 – 30	9
30 – 40	12
40 – 50	7
50 – 60	9

CHAPTER II

MEASURES OF DISPERSION

In the previous chapter, we have studied that the values in a series will have a tendency to cluster around certain values and those central values are known as measures of locaticn or central tendencies. We have also studied that different measures will be very helpful for comparing different disiributions. In this chapter, we shall study some other measures which are just opposite to the measures of location or measures of concentration. These new measures therefore indicate the variation or dispersion. Hence, these new measures are called Measures of dispersion.

Before we proceed further, let us consider the business of three merchants dealing in perishable articles like vegetables. We shall examine the business from the sales and decide the utility of continuing the business.

	Merchant A Sales Rs.	Merchant B Sales Rs.	Merchant C Sales Rs.
Monday	40	35	40
Tuesday	75	50	40
Wednesday	25	45	40
Thursday	80	40	40
Friday	50	30	40
Saturday	30	40	40
Total sales	300	240	240

Average sales per day for A = 300/6 = Rs. 50.

„ for B ⟶ = 240/6 = Rs. 40.

„ for C ⟶ = 240/6 = Rs. 40.

The total weekly sales of A is Rs. 300 while that of the merchants B and C is Rs. 240. Hence we naturally decide that 'A' is getting better business than B and C. We can arrive at the very same conclusion by comparing the average daily sales instead of comparing the total sales for 6 days. The average sale of 'A' comes to Rs. 50 per day while that of 'B' and 'C' is only Rs. 40. Hence the business of 'A' is better than B and C. A person whc is not prudent enough may come to the above conclusion.

Let us now examine the trend of business an each day. In the case of A we find a wide fluctuation ranging from Rs. 25 to Rs. 80, while in the case of B it varies from Rs. 30 to Rs. 50. But in the case of 'C' there is no variation at all. The variation in the case of 'B' is not so wide as in the case of 'A'. The gap between the highest sales and the lowest sales in the case of A is Rs. 55 (80—25). While in the case of B, the gap is only Rs. 20 (50—30). which is about 50% of his average sales. In the case of 'C' it is '0'. Because of this we may naturally comment that the business of B is mcre or less more steady than that of A. In the case of 'C' it is the most steady. In the case of A, the sales on Tuesday reaches Rs. 75. Expecting the same type of high sales on the next day, if he purchases greater quantity for the next day, he will sustain a heavy loss since the sales comes to Rs. 25 resulting in heavy stock of perishable goods. If he reduces the purchase on the next day because of the poor sales on the previous day, he will have a good demand cn the next day. Unless he has sufficient stcck, he will be losing the customers. Business having such vagaries in the sales is really not a good. Though the sales of A is better from the point of view of average sales, the sales of 'C' is better from the aspect of steadiness.

From the above observation, we may infer that we should not decide the efficiency of the business either by the total sales or by the average sales, but we must also take into consideration the

variation or change in the day to day sales. As we have said that the values have a tendency to centre around a value we can also say that the values have a tendency to disperse or deviate or vary from the central values. Similar to the measures of central tendency we can have measures of variations or Measures of dispersion or Measures of deviation.

There are different measures of dispersion. They are (1) Range, (2) Mean Deviation, (3) Standard Deviation, (4) Variance, (5) Co-efficient of variation and (6) Semi-Inter quartile deviation.

1. RANGE

Range is the simplest measure of dispersion. It is defined as the difference between the highest and the lowest or the maximum and the minimum or the largest and the smallest values in the series or distribution. It is also expressed in the same unit as the original values.

Example (1): Weight of persons in a factory in kg.

50, 60, 52, 45, 49, 35, 42, 40.

Maximum value = 60 kg.

Minimum value = 35 kg.

Range = 60—35 = 25 kg.

Example (2): We shall consider another series of persons whose weights are given in terms of some other units (say lb.)

120, 130, 125, 160, 112, 115, 140, 105.

Maximum value = 160 lb.

Minimum value = 115 lb.

Range = 160—115 = 45 lb.

In the above two cases, the range is 25 kg. in one series and in the other it is 45 lb. It is very difficult to compare the two distributions with the help of Ranges, when the values of the distributions are expressed in different units, here, kg. and lb. From the absolute values and from the actual numerical values without unit of measurement we will say that the second series is having the greater difference 45 than the first which is having only 25. While in the case of measures of central tendencies, the greater the values, the greater the importance. But in the case of dispersion, the lesser the value the greater the importance. From this angle, we may say that series 1 is better which may not be correct.

For easy comparison, the units of the values of different distribution should be the same. This can be seen from the following situation

Suppose we want to study the income of industrial workers in different countries like Great Britain, Germany, Japan, Russia, United States and India. The income of the workers in these countries will be expressed in terms of their currency. Let us say that the average income and range in their income are as follows

	Country	Average monthly income	Range in the income
1.	Great Britain	300 pounds	75 pounds
2.	Germany	500 Mark	45 Mark
3.	Japan	450 Yen	50 Yen
4.	Russia	350 Roubles	35 Roubles
5.	United States	700 Dollars	100 Dollars
6.	India	400 Rupees	80 Rupees

The comparison is difficult. The difficulty is due to the different units. For the sake of comparisons we cannot do away with the units. Therefore, we have to think of an alternative method for comparision where the units do not play a part. We should have *relative measures*.

Co-efficient of Dispersion (or) Co-efficient of Range

A relative measure is not expressed in any unit. It is free from units of measurements. It is only a mere number. The relative measure is called Co-efficient of dispersion. It is defined as follows:

$$\text{Co-efficient of dispersion} = \frac{\text{Difference between the largest and the smallest values}}{\text{Sum of the largest and the smallest values.}}$$

If the largest value is denoted by 'L' and the smallest value by 'S' we can have the formula for defining Co-efficient of dispersion as follows:

$$\text{Co-efficient of dispersion} = \frac{L - S}{L + S}$$

We shall calculate the co-efficient of dispersion for the two examples considered earlier.

		Example I	Example II
Largest value		60	160
Smallest value		35	115
	Sum	95	275
	Difference	25	45

Co-efficient of dispersion

(Example I) $= \dfrac{L-S}{L+S} = \dfrac{60-35}{60+35} = \dfrac{25}{95}$

$= 0.26$

Co-efficient of dispersion

(Example II) $= \dfrac{160-115}{160+115} = \dfrac{45}{275} = \mathbf{0.16.}$

We find that the second example has a lesser value of co-efficient of dispersion. Hence the second is better.

(1) Calculate the co-efficient of dispersion.

30, 45, 50, 70, 75.

Largest value 'L' = 75.

Smallest value 'S' = 30.

Co-efficient of dispersion $= \dfrac{L-S}{L+S} = \dfrac{75-30}{75+30} = \dfrac{45}{105}$

$= \mathbf{0.43}$

(2) Calculate the Range and the co-efficient of dispersion in the following distribution:

Weight (lb.)	No. of students
40 – 50	4
50 – 60	2
60 — 70	5
70 — 80	2
80 — 90	1

Highest value $= L = 90.$

Smallest valu $= S = 40.$

$\therefore$ Range $= L - S = 90 - 40 = 50$ lb.

$$\text{Co-efficient of Range} = \frac{L - S}{L + S} = \frac{90 - 40}{90 + 40} = \frac{50}{130} = 0.38$$

Merits and Demerits of Range

Range is easy for calculation and understanding. But it has its own defects. Range gives the difference between the highest and the lowest values of the variable. Therefore, we consider only two values namely the highest value and the lowest value and from these two values we come to a conclusion about the dependability of the distribution. But the distribution may have a number of values or a number of units and expressing an opinion about all the members in the distribution based on the observation of only two members may not be a sound proposition. Whatever we say it should be based on the observations of the values of all the members in the group and not on the basis of only two members. Of course it may be a quick process but it has its own defects.

Defects

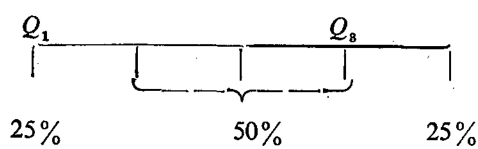
1. There are seldom highest and lowest values.
2. The occurence of one of the values either highest or lowest has considerable effect on the value of the Range.
3. It is not representative of all values in the series.

Therefore, we have to think of some other alternative measures of deviation which will throw light on the deviatinos of all the

values or the values of all the members in the group or it should give an overall picture of all the members or it should be a true representative of all the members in the group.

2. QUARTILE DEVIATION — Q

In the previous chapter dealing with measurement of central tendency we have studied two quartiles namely Q_3 upper quartile and Q_1 the lower quartile. Q_3 is that value of the variable which divides the distribution into two parts such that 25% of the total number of units will have value greater than Q_3 and the remaining 75% of the units will have value less than Q_3. Q_1 is just opposite to this. Q_1 is that value of the variable which divides the population into two parts such that 25% of the total number of members will have value less than Q_1 and the remaining 75% of the members will have value greater than Q_1. It can be further explained. In the case of Q_3, 25% of the members will have value greater than Q_3 and in the case of Q_1, 25% of the members will have values less than Q_1. Therefore the range between $Q_3 - Q_1$ will contain 50% of the population units. This range is called the Interquartile Range. If this range is divided by 2, we will have quartile deviation, Q, which is otherwise called as Semi-Interquartile Range.



Hence we can have the Semi-Inter Quartile Range as a measure of deviation and it is denoted by the letter 'Q'.

$$Q = \frac{Q_3 - Q_1}{2}$$

'Q' is also expressed in the same units as the original values of the distribution. Hence the comparison of the two or more distributions, each in different units of measures is difficult.

Hence, we will consider here the computation of co-efficient of Quartile dispersion free from the values of the quartiles themselves.

Calculate (1) Quartile Deviation.

(2) Co-efficient of Quartile dispersion.

$Q_1 = 45$ kg. $Q_3 = 75$ kg.

1. Quartile Deviation $= \dfrac{Q_3 - Q_1}{2} = \dfrac{75 - 45}{2} = \dfrac{30}{2}$

$= 15$ kg.

2. Co-efficient of quartile dispersion $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

$$= \frac{75 - 45}{75 + 45}$$

$$= \frac{30}{120} = 0.25$$

Co-efficient of quartile dispersion is free from the units of measurements. It is expressed only in terms of numerical values. Hence easy for comparison. The computation is easy when the values of quartiles are given. But the computation is difficult if we have to calculate it after computing the values of quartiles from the original values. This is calculated with reference to the position of the items and not with reference to the values of the items.

3. MEAN DEVIATION (M.D)

When we studied the Ranges, we have seen that Range is based on only two values ie: the maximum and the minimum values. We have also seen that Quartile deviation is influenced by the positions of the items and not by the values of items. Now we shall see measures which are influenced by the values of the items and that too by the values of all the items in the series or distribution. One such measure is Mean deviation (M.D.) or Average Deviation.

Mean Deviation (M.D.)

For a given set of data we should calculate the average or Arithmetic Mean $\bar{x}$. We should then subtract each value from the Mean and find out the difference (d) or deviation of each value from the Mean. Next we should find the sum of the total deviations of all values from Mean $\Sigma\ d$. The sum of total deviations should be divided by the number of items 'n'. $\frac{\Sigma\ d}{n}$ This is called Average or Mean deviation per item.

Let us consider the following example:

	Sales Shop I Rs.	Sales Shop II Rs.
Monday	40	35
Tuesday	75	50
Wednesday	25	45
Thursday	80	40
Friday	50	30
Saturday	30	40
	300	240

Average = 300/6 = Rs. 50/- 240/6 = Rs.40/-

The daily sales in the case of two shops are given above. We find that the average sales in the shops is Rs. 50/- and Rs. 40/- per day. Let us now calculate the difference of the sales of each day from the average sales.

	Shop I	Shop II
Monday	40 — 50 = — 10	35 — 40 = — 5
Tuesday	75 — 50 = + 25	50 — 40 = 10
Wednesday	25 — 50 = — 25	45 — 40 = + 5
Thursday	80 — 50 = + 30	40 — 40 = 0
Friday	50 — 50 = 0	30 — 40 = — 10
Saturday	30 — 50 = — 20	40 - 40 = 0
Total Difference	0	0

Average difference = 0/6 = 0.   Average difference = 0/6. = 0.

We find that the total difference is 0 in both the cases. Consequently, the average difference will also be '0'. Not only in these two cases, but also in all cases, total sum of deviations will be equal to 0, if the deviations are calculated from the Arithmetic Mean. This is one of the important properties of the Arithmetic Mean, which we have studied in the previous chapter. Some of the deviations are positive since the values are greater than the Mean and the rest of the differences are negative since the values are smaller than the Mean. The sum of all positive deviations will be equal to the sum of all negative deviations and they get cancelled mutually. Thus the total of deviations is equal to '0'.

In order to have effective comparison we should overcome the above difficulty. The total deviation comes to '0' because of the occurrence of positive and negative values of deviations. We can ignore the signs of the deviations. In other words, we

should consider only the numerical values without their signs or we should consider all the values as positive figures. Consideration of only the value is known as consideration of only absolute values. The sum of the absolute values divided by the number of units will give the average deviation which is called Mean Deviation.

Absolute Deviation *in the previous examples*

Days	Case (*I*)	Case (*II*)
Monday	10	5
Tuesday	25	10
Wednesday	25	5
Thursday	30	0
Friday	0	10
Saturday	20	0
Total	110	30

Mean deviation $= \frac{110}{6} = 18\ 1/3.$ $30/6 = 5.$

The Mean Deviation of *I* is greater than the *M.D.* of *II*. Hence, *II* is more dependable. The lesser the value of the Mean Deviation the greater the dependability.

$$\text{Mean Deviation} = \frac{\Sigma \mid \text{Total deviation} \mid}{n}$$

The two lines on either side indicate that the figures are only absolute values without any consideration of the sign.

$$\text{Mean Deviation} = \frac{\Sigma \mid (x - \bar{x}) \mid}{n}$$

In case we get frequencies the formula can be suitably modified for the multiplication of the difference by the respective frequency.

$$\frac{\Sigma\,(x_i - \bar{x})\,f_i}{N}$$

$$\text{where } N = \Sigma\, f_i$$

Defects

Ignoring the sign of the deviation may not be a good proposition since we make the differences as artificial.

The peculiar situation of getting the total deviation as 0, or taking only the absolute deviation thereby considering only the artificial differences can be avoided if we calculate the difference of the values from any other value instead of from Arithmetic Mean. We can calculate the difference from the Monday value or Tuesday value. But each person will select different values for comparison and therefore there may not be any uniformity in the choice of the base. Unless there is uniformity in the choice of the base among the examiners, we cannot have effective comparison and the comparison will then be meaningless.

We can calculate the Mean deviation not only from Mean but also from Median or Mode. But calculation of Mean deviation from Median or Mode will involve considerable computation since the calculation of Median and Mode themselves are rather more cumbersome than the calculation of Mean. Therefore, calculation of Mean deviation from Mean is more easy and hence widely used.

Mean co-efficient of dispersion

As we have calculated relative measures of dispersions from Range and Quartile deviation, we can also calculate relative measure of dispersion from Mean Deviation also. This is called Mean Co-efficient of dispersion. This can be calculated as follows:

$$\text{Mean co- efficient of dispersion} = \frac{\text{Mean Deviation from Mean}}{\text{Mean}}$$

$$= \frac{\Sigma\,|\,x - \bar{x}\,|}{\bar{x}}$$

Mean co-efficient of dispersion will be expressed in mere number—without any unit of measurement. Therefore, it helps for easy comparison. In the previous example, Mean and Mean Deviation are as follows:

	Case I	Case II
Mean $\bar{x}$	50	40
Mean Deviation	110/6	30/6

$$\text{Mean co-efficient of dispersion} = \frac{110}{6 \times 50} \qquad \frac{30}{6} \times \frac{1}{40}$$

$$= 11/30 \qquad = 1/8.$$

4. VARIANCE (V) = σ^2

We have seen in the calculation of Mean Deviation that uniform base has to be adopted. If not, each one would calculate the Mean Deviation from different base either from Mean, or from Median or from Mode or from any other arbitrary value as one desires. For the sake of uniformity the Mean deviation is calculated from Mean.

We have seen that when deviation of each value is calculated from the Mean, the total sum of deviations is equal to 0. In order to overcome this situation, we have suggested that absolute values of the deviations ie: the numerical values of the deviations without the sign can be adopted. But it does not seem to be quite convincing. It appears to be an artificial way of overcoming the difficulty.

We can overcome the difficulty of the sign of the deviations by another way instead of ignoring them. In this process we can square the deviation. The square of the deviation will always be

positive irrespective of the fact whether the sign of the deviation is positive or negative. Since the square of a negative quantity will always be a positive quantity. We can find the total sum of squares of deviations. The total sum of squares of deviations can be divided by the number of items (n) and the average can be calculated. This is called Average Square of Deviations or Mean Square Deviation. It is known as Variance (v) in statistics. We can examine this with the help of the example given below:

		Sales Rs.	Case I (deviation) $d = x - \bar{x}$	Square of the deviation $d^2 = (x - \bar{x})^2$
1.	Monday	40	$40 - 50 = -10$	100
2.	Tuesday	75	$75 - 50 = 25$	625
3.	Wednesday	25	$25 - 50 = -25$	625
4.	Thursday	80	$80 - 50 = +30$	900
5.	Friday	50	$50 - 50 = 0$	–
6.	Saturday	30	$30 - 50 = -20$	400
	Total	300		2650

Mean $= 300/6 =$ Rs. 50/-

$$\text{Mean Square Deviation } (v) = \frac{2650}{6} = 441.7 \text{ per day.}$$

Let us examine the Case II:

		Sales Rs.	Deviation $d = x - \bar{x}$	$d^2 = (x - \bar{x})^2$
1.	Monday	35	-5	25
2.	Tuesday	50	10	100
3.	Wednesday	45	5	25
4.	Thursday	40	0	—
5.	Friday	30	-10	100
6.	Saturday	40	0	—
	Total	240		250

Mean $= 240/6 = 40$. Mean Square Deviation $= \dfrac{250}{6} = 41\dfrac{2}{3}$ per day.

We find that the value of Mean Square Deviation or Variance in the second case is smaller than that in the first case. Consequently, we can say that the sales in the second case are more steady and reliable.

5. STANDARD DEVIATION — σ (Sigma)

Standard Deviation is another measure of dispersion called sigma denoted by the Greek letter (small) σ. This is the most important measure or parameter in statistics and is widely used in all statistical applications. Hence greater care on the part of the students is required in the study and computation of standard

deviation. It can be rightly said that the science of statistics is revolving with Arithmetic Mean as the centre and standard deviation as the radius.

For calculating the variance (v) or Mean Square Deviation (M.S.D.) we have followed the following procedurs:

(1) We have found out the sum of the values of the given item by adding all the values Σx.

(2) We have calculated the Arithmetic Mean of the given values by dividing the total value by the number of items: $\frac{\Sigma x}{n} = \bar{x}$

(3) We have subtracted the Arithmetic Mean from each of the given values and found the difference or deviation for each value. $(x - \bar{x}) = d$.

(4) Then we have squared each of the deviations (d^2).

(5) We have found out the sum of squares of all deviations by totalling them $\Sigma d^2 = \Sigma (x - \bar{x}^2)$

(6) The sum of the squares cf deviations thus arrived at was divided by the number of items to find out the Mean Square of Deviation. $\frac{\Sigma d^2}{n} = \frac{\Sigma (x - \bar{x})^2}{N}$

(7) This Mean Square deviation is nothing but the variance.

(8) The Square Root of the Mean Square Deviation is called Standard Deviation denoted by the letter σ.

$$\sigma = \sqrt{v} = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$$

In order to overcome the difficulty encountered due to the sign in the case of individual deviation we have squared them and finally arrived at this, since what we want is the average deviation and not average square deviation. Hence we have to find out the square root of the variance since we have originally squared

the individual deviation. There will be 2 values for the square root, one positive and another negative. We would take the positive value.

We shall examine the same examples considered previously.

Day	Sales in Rs. x	Deviation from Mean $x - \bar{x}$	Square of the deviation $(x - \bar{x})^2$
1. Monday	40	– 10	100
2. Tuesday	75	25	625
3. Wednesday	25	– 25	625
4. Thursday	80	30	900
5. Friday	50	0	–
6. Saturday	30	—20	400
Total	300		2650

Mean = 300/6 = Rs. 50 per day.

Sum of square of deviation = 2650; $n = 6$.

Mean Square Deviation = 2650/6
= 441·7

Standard Deviation = $\sqrt{441 \cdot 7}$
= 21 rupees per day.

It should be noted that the Standard Deviation will always be expressed in the same unit as the original items are expressed.

Root Mean Square Deviation

The Standard Deviation is otherwise known as Root Mean Square Deviation. If we read this name from right to left (Deviation, Square, Mean, Root) instead from left to right, it would indicate the processes involved in the computation of Standard Deviation.

1.	Deviation	::	Find out the deviation of each value from the Mean. d
2.	Square	::	Square each deviation. d^2
3.	Mean	::	Find the Mean Square of the deviation. d^2/n
4.	Root	::	Find out the Square root of the Mean Square of the deviation. $\sqrt{\frac{d^2}{n}}$

The lesser the value of the Standard deviation the greater the reliability of the values of the distribution. In other words, it will indicate that each value does not vary or differ very much from the other. If the standard deviation is 0, it will indicate all the values are same.

In the same way, we can calculate Root Mean Square Deviation from any value other than Mean. But the Mean Square deviation calculated from the Mean is taken as a standard. Hence it is known as Standard Deviation.

Co-efficient of Variation (C.V.)

Before we study more about Standard Deviation we should study another measurement called the co-efficient of variation.

We have seen that the standard deviation is always expressed in the same unit of measures as the original item. Therefore, different distributions having values in different units of measurement will have standard deviations also in different units of measurement. In such situation the comparision will not be effective or rather possible. If we want to have effective comparison,

with the help of standard deviation, we should get rid of the unit of measurement. This is not possible. We can overcome this by finding out a relative measurement free from units of measurement. For this purpose, the standard deviation can be divided by the Arithmetic Mean $\frac{\sigma}{\bar{x}}$

$= \frac{\text{Standard deviation}}{\text{Arithmetic Mean}}$. This ratio will be free from units of measurement and will be a mere number. The ratio will be too small. Hence it is multiplied by 100.

$$\frac{\sigma}{\bar{x}} \times 100.$$

In this process, the standard deviation is expressed as a percentage of the Mean. This value is known as co-efficient of variation and represented by the letter '*C.V*'.

$$C.V = \frac{\sigma}{\bar{x}} \times 100$$

In the above example, the Arithmetic Mean is Rs. 50/- and the standard deviation is Rs. 21.

$$\therefore \text{ Co-efficient of variation} = \frac{21}{50} \times 100$$

$$= \mathbf{42} \text{ (mere number)}$$

Advantages in calculating the Mean Square Deviation from Arithmetic Mean

One great advantage in calculating the variance and standard deviation from the Mean is that they will be minimum in their values. If we calculate the variance or deviation from any other value, which is either greater or less than the Mean, the variance will be greater in value than the one computed from the Mean. This gives another important property of the Arithmetic Mean namely that the sum of the squares of diviations taken from the Arithmetic Mean will always be the minimum.

The properties of the Arithmetic Mean can be summarised as follows:

1. The sum of the deviations taken from the Mean will always be 0.
2. The sum of the Squares of deviations taken from the mean will always be the minimum.

Whatever apply to the sum of the deviations or sum of the squares of deviations will apply to the average deviation or Mean deviation or Mean Square deviation. We have already verified the first property. Now we shall verify the second property.

We shall calculate the Variance from another arbitrary value say 'A' instead of the Arithmetic Mean.

In this context it is better to refresh our mind. We know that $(a+b)^2 = a^2 + 2ab + b^2$. This formula is now applied here. The deviation of each value from one arbitrary value 'A' can be written as the sum of deviation of that value from the Mean and the difference between the mean and the arbitrary value say A.

'x' $- A$ = Deviation of the value from the arbitrary value.

$x - \bar{x}$ = Deviation of the value from Mean.

$\bar{x} - A$ = Difference between the Mean and the Arbitrary value.

$$x - A = x - \bar{x} + \bar{x} - A,$$

$$= (x - \bar{x}) + (\bar{x} - A)$$

$$(x - A)^2 = \{(x - \bar{x}) + (\bar{x} - A)\}^2$$

$$= (x - \bar{x})^2 + 2\,(x - \bar{x})(\bar{x} - A) + (\bar{x} - A)^2$$

This is the expansion we get for one value. But there are many items and we have to calculate the square of the deviation for each and every item. The sum of squares of deviations can

be calculated by adding the squares of differences of all the items. This can be written as follows:

$$\underset{(1)}{\Sigma (x - A)^2} = \underset{(2)}{\Sigma (x - \bar{x})^2} + \underset{(3)}{\Sigma 2(x - \bar{x})(\bar{x} - A)} + \underset{(4)}{\Sigma (\bar{x} - A)^2}$$

(1) Sum of the squares of the difference between the values and the arbitrary value.

(2) Sum of the squares of the deviations of the values from the Mean.

(3) Twice the sum of the products of the difference $(x - \bar{x})$ and $(\bar{x} - A)$.

(4) Sum of the squares of differences of the Mean and the arbitrary value A.

Let us take the term (3).

$\Sigma 2 (x - \bar{x}) (\bar{x} - A)$

2 is a constant number and hence can be taken outside the Σ symbol.

$\bar{x}$ is a constant as far as a particular distribution is concerned

A is also constant number.

Hence $(\bar{x} - A)$ is a constant number and it can also be taken outside Σ.

So we can take out $2 (\bar{x} - A)$ outside the sigma symbol.

$2 (\bar{x} - A) \Sigma (x - \bar{x})$

But $\Sigma (x - \bar{x})$ = the sum of the deviations taken from the Arithmetic Mean, which is equal to '0'.

Hence $\Sigma (x - A)^2 = \Sigma (x - \bar{x})^2 + 0 + \Sigma (\bar{x} - A)^2$

$$\Sigma (x - A)^2 = \Sigma (x - \bar{x})^2 + N \cdot (\bar{x} - A)^2$$

Sum of the squares of deviations from A =Sum of the Squares of deviations taken from Mean + Sum of the squares of the differences between Mean and A.

Let us divide all the terms by 'N'.

$$\frac{\Sigma (x - A)^2}{N} = \frac{\Sigma(x - \bar{x})^2}{N} + \frac{\Sigma (\bar{x} - A)^2}{N}$$

Variance about A = Variance about $\bar{x}$ + Square of the difference between $\bar{x}$ and A.

Variance about A = Variance + Square of the difference between $\bar{x}$ and A.

Since $\frac{\Sigma (\bar{x} - A)^2}{N} = \frac{N (\bar{x} - A)^2}{N} = (\bar{x} - A)^2$

Let us denote the various terms as follows:

Variance about any arbitrary value (A) $= S^2$

Variance about Mean $\bar{x}$ $= V$

Difference between A and $\bar{x}$ $= d$

$$\therefore\ S^2 = V + d^2 \text{ or } V = S^2 - d^2$$

ie: $\sigma^2 = S^2 - d^2$

S^2 is minimum when $d = O$ ie when $A = \bar{x}$. The minimum value of $S^2 = \sigma^2$.

So it is seen from the above that V, ie: the Mean Square Deviation taken from Arithmetic Mean will always be the minimum. When the Mean Square of the deviation is minimum, the sum of the square of the deviation will also be the minimum if it is calculated from the Arithmetic Mean. It may be noted that $d = (\bar{x} - A)$ will be positive when A is less than $\bar{x}$. $d = (\bar{x} - A)$ will be negative when A is greater than $\bar{x}$. However, d^2 will be always positive since the square of any number, either positive or negative, will always be a positive.

Advantages of Standard Deviation

It is based on all observations. It can be easily calculated, easily understood. It is amenable to algebraic treatment.

I. Direct Method

$$V = \frac{\Sigma (x - \bar{x})^2}{N}$$

$x - \bar{x}$ = Deviation.

In order to find out the variance, we have to calculate the average $\bar{x}$. Then the deviation of each of the values from the Mean has to be squared $(x - \bar{x}^2)$. Then we have to add the squares of each deviation and find out the sum of squares of the deviation. $\Sigma (x - \bar{x})^2$.

This sum of squares of the deviations has to be divided by 'N' to get the Mean Square Deviation.

The various processes involved can be avoided and we can calculate the V from the original values themselves.

We know that

$$V = \frac{\Sigma (x - \bar{x})^2}{N}$$

We know that $(a - b)^2 = a^2 - 2ab + b^2$

$\therefore (x - \bar{x})^2 = x^2 - 2x\bar{x} + \bar{x}^2$

When we take the sum of squares of deviations

$$\Sigma (x - \bar{x})^2 = \Sigma x^2 - \Sigma 2 x \bar{x} + \Sigma \bar{x}^2$$

$$\Sigma (x - \bar{x})^2 = \Sigma x^2 - 2\bar{x} \Sigma x + \Sigma \bar{x}^2$$

(Since 2 is a constant number and $\bar{x}$ is a constant value as far as a particular distrbution is concerned, 2 $\bar{x}$ can be taken outside the sigma symbol).

$$= \Sigma x^2 - 2 \bar{x} \Sigma x + N \bar{x}^2$$

But $\Sigma x = N \bar{x}$

$= \Sigma x^2 - 2N \bar{x}^2 + N \bar{x}^2$ (since $\frac{\Sigma x}{N} = \bar{x} \therefore \Sigma x = N\bar{x}$

$= x^2 - N \bar{x}^2$

Let us divide it by N to get V,

$$\frac{\Sigma (x - \bar{x})^2}{N} = \frac{\Sigma x^2}{N} - \frac{N \bar{x}^2}{N}$$

$$= \frac{x^2}{N} - \bar{x}^2$$

Mean square of the original value — Square of the Mean.

This can be further simplified.

(1) $= \dfrac{\Sigma x^2}{N} - \left\{ \dfrac{\Sigma x}{N} \right\}^2$ (Since $\bar{x} = \Sigma x/N$)

(2) $= \dfrac{\Sigma x^2}{N} - \dfrac{(\Sigma x)^2}{N^2}$

(3) $= 1/N \left\{ \Sigma x^2 - \dfrac{(\Sigma x)^2}{N} \right\}$

The students should clearly notice the difference between Σx^2 and $(\Sigma x)^2$

Σx^2 = Sum of the squares of each value.

$(\Sigma x)^2$ = Square of the sum of all the values.

Suppose there are two items; 3 and 4

$\Sigma x^2 = 3^2+4^2 = 9+16 = 25.$

$(\Sigma x)^2 = (3+4)^2 = 7^2 = 49.$

Any one of the three formuale can be followed.

Example Find the standard deviation of the following series:

Weight of bags (kg)	x^2
67	4489
75	5625
80	6400
83	6889
85	7225
93	8649
97	9409
91	8281
98	9604
769	66571

(1) $\Sigma \dfrac{x^2}{N} - \bar{x}^2$

$$= \frac{66571}{9} - (85.44)^2$$

$$= 7396.78 - 7299.99$$

$$= 96.79$$

$$\sigma = \sqrt{96.79}$$

$$= 9.8$$

2nd Method:

$$\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

$$= 7396.78 - (769/9)^2$$

$$= 7396.78 - \frac{591361}{81}$$

$$= 7396.78 - 7300.75$$

$$= \sqrt{96.03}$$

$$= 9.8$$

3rd Method:

$$1/N \left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{N} \right\}$$

$$= 1/9 \left\{ 66571 - \frac{769 \times 769}{9} \right\}$$

$$= 1/9 \left\{ 66571 - \frac{591361}{9} \right\}$$

$$= 1/9 \ (66571 - 65707)$$

$$= 1/9 \times 864 = \frac{864}{9} = 96.$$

$$= \sqrt{96}$$

$$= 9.8$$

All the three formulae are one and the same. The first is the simplest and easy for remembrance. The second and third are nothing but modifications of the first formula. Therefore, we shall follow the first in our further calculations.

In the above example, we have worked out the variance and the standard deviation from ungrouped data by shortcut method. The formula used is:

$$V = \frac{\Sigma x^3}{N} - \bar{x}^2$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N} - \bar{x}^2} = \sqrt{v}$$

This is the formula for the direct method.

Example 2: Calculate variance and Standard Deviation for the values.

30, 80, 60, 70, 20, 40, 50.

Values x (Rs.)	x^2
30	900
80	6400
60	3600
70	4900
20	400
40	1600
50	2500
350	20300

$n = 7.$

$\Sigma x = 350.$

$$\bar{x} = \frac{350}{7} = 50.$$

$\bar{x}^2 = 50 \times 50 = 2500$

$\Sigma x^2 = 20300$

$$\text{Variance} = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$$\frac{\Sigma x^2}{n} = \frac{20300}{7}$$

$$V = \text{Variance} = \frac{20300}{7} - 2500$$

$$= 400$$

σ : Standard Deviation $= \sigma = \sqrt{400}$

$= \text{Rs. } 20$ per item.

Shortcut Method

We shall adopt shortcut method for calculation of standard diviation. In this shortcut method we calculate deviation from an assumed mean. The following steps are adopted.

1. Let us first assume an arbitrary value, say 60 in this case as A.
2. Calculate the deviation of each value from this assumed mean and let it be denoted by the letters $(d) = x - A$.
3. Square each such deviation: $d^2 = (x-A)^2$.
4. Find the sum of squares of deviation $= \Sigma d^2$
5. Then calculate the standard deviation of 'd' with the help of the following formula

$$\sigma^2 = \frac{\Sigma (x - A)_2}{n} - \left\{\frac{(\Sigma x - A)}{n}\right\}^2$$

$$= \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2$$

$$\sigma^2 = \frac{\Sigma d^2}{n} - \bar{d}^2$$

when $A = 0$ the formula will become

$$\sigma^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

	x	$x - A = d$	d^2
	30	—30	900
	80	20	400
A—	60	0	0
	70	10	100
	20	—40	1600
	40	—20	400
	50	—10	100
		Σd — 70	3500

$$\bar{d} = \frac{-70}{7} = -10.$$

$$\bar{d}^2 = -10 \times -10 = 100.$$

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma d^2}{n} - \bar{d}^2}$$

$$= \sqrt{\frac{3500}{7} - 100}$$

$$= \sqrt{500 - 100}$$

$$= \sqrt{400}$$

$$= \text{Rs. } 20 \text{ per head.}$$

The standard deviaton of 'd' is the same as the standard deviation of x.

Example 3

Calculate the standard deviation by the shortcut method:

Weight of Bags x	$d = x - A$	d^2
67	-16	256
75	-8	64
80	-3	9
85	2	4
83—A	0	0
93	10	100
97	14	196
91	8	64
98	15	225
	22	918

Let us assume 83 as A. and calculate the deviation of each value from 83.

$$n = 9; \quad \Sigma d = 22; \quad \bar{d} = 22/9 = 2.44 \quad \bar{d}^2 = 5.95$$
$$\Sigma d^2 = 918.$$

Standard Deviation of 'd' $= \sqrt{\dfrac{\Sigma d^2}{n} - \bar{d}^2}$

$$= \sqrt{\frac{918}{9} - 5.95} = \sqrt{96.05}$$
$$= \mathbf{9.8}$$

This is the value that we have obtained in our previous example No. (1) under Direct Method.

II. Calculation of Standard Deviation from Discrete frequency distribution

A. Direct Method

We can calculate the standard deviation by the direct method as follows:

1. We should first find out the Arithmetic Mean of the distribution by multiplying each value by its respective frequency and dividing the total of the products by the sum of the total frequencies. The following formula can be used:

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

2. Deviation of each value from the Mean can be calculated $(x - \bar{x})$.

3. Square of each deviation should be calculated $(x - \bar{x})^2$

4. Square of each deviation should be multiplied by the respective frequency $(x - \bar{x})^2 f$.

5. Sum of the squares of deviation multiplied by the frequencies should be calculated by adding $\Sigma (x - \bar{x})^2 f$.

6. The average of the squares of deviation should be calculated by dividing the sum of squares of deviation by the sum of frequencies.

$$\frac{\Sigma (x - \bar{x})^2 f}{\Sigma f}$$

This is called Mean Square Deviation or Variance.

7. We should calculate the Square root for this Mean Square Deviation and this is the standard deviation.

$$\sqrt{\frac{\Sigma (x - \bar{x})^2 f}{\Sigma f}}$$

The method is more or less the same as the method adopted for ungrouped data. The only difference is that we have used the frequency in two stages as follows:

(1) We have multiplied the value by the frequency for calculating the Mean;

(2) We have multiplied the square of the deviation by the frequency.

Example 4: Calculate the Standard Deviation for the following distribution:

Value (x)	Frequency f	Deviation xf
30	3	90
80	5	400
60	6	360
70	10	700
20	3	60
40	2	80
50	1	50
	30	1740

$N = \Sigma f = 30. \quad \Sigma f.x = 1740.$

$$\therefore \quad \bar{x} = \frac{1740}{30} = 58$$

x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2.f$
30	3	30 — 58 = —28	784	2352
80	5	80 — 58 = 22	484	2420
60	6	60 — 58 = 2	4	24
70	10	70 — 58 = 12	144	1440
20	3	20 — 58 = —38	1444	4332
40	2	40 — 58 = —18	324	648
50	1	50 — 58 = — 8	64	64
	30			11280

$$V = \frac{\Sigma (x - \bar{x})^2.f}{\Sigma f} = \frac{11280}{30} = 376$$

$\sigma = \sqrt{376}$ = Rs. 19.4 per head.

B. Shortcut Method

We can calculate the Standard Deviation by shortcut method. We can use an assumed Mean and find out the standard deviation. The various processes are the same as before.

Let us take 50 as the assumed Mean denoted by A. Each deviation will be $x - A$ and its square will be $(x - A)^2$. The

process is given below:

x	f	$(x-50)$ d	$(x-50)f$ fd	$(x-50)^2$ d^2	$f(x-50)^2$ $f.d^2$
30	3	—20	—60	400	1200
80	5	30	150	900	4500
60	6	10	60	100	600
70	10	20	200	400	4000
20	3	—30	—90	900	2700
40	2	—10	—20	100	200
50	1	0	0	0	0
			240		13200

$N = \Sigma f = 30.$ $\quad\Sigma f d = 240$

$$\therefore \quad \bar{d} = \frac{240}{30} = 8. \qquad \bar{d}^2 = 64.$$

$\Sigma f d^2 = 13200$

$$\text{Mean Square Deviation} = \frac{13200}{30} - 64.$$

$$= 440 - 64 = 376.$$

$$\text{Standard Deviation} = \sqrt{376} = \text{Rs. } 19.4 \text{ per head.}$$

In this process we find that the Standard Deviation of the original value and the Standard Deviation of the new variable obtained by substituting $x - A$ are one and the same. In this process, the difficulty involved in squaring the deviation is greatly reduced.

III. Calculation of Standard Deviation from continuous frequency distribution

A. Direct Method

The various steps involved are as follows:

(1) Mid value of each class should be calculated and afterwards the methods are the same as in the case of discrete frequency distribution.

(2) The mid-value has to be multiplied by the respective frequency and fromth is total, the Arithmetic Mean has to be calculated by using the formula

$\bar{x} = \frac{\Sigma x_i f_i}{\Sigma fi}$ where x_i indicates the mid-value.

(3) The deviation of each value has to be calculated from the Mean $= x - \bar{x} = \bar{d}$.

(4) The deviation obtained has to be squared $(x - \bar{x})^2$.

(5) Each square of the deviation has to be multiplied by the respective frequency $(x - \bar{x})^2 f$.

(6) From the total square of the deviation, we have to find out the Mean Square Deviation by dividing it by $\Sigma f = N$.

$$= \frac{\Sigma (x - \bar{x})^2 f}{\Sigma f}$$

(7) The square root of the Mean Square deviation will be the Root Mean Square deviation:

$$\sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2 f}{\Sigma f}}$$

Example 5: Calculate the standard deviation for the following frequency distribution.

Classes Weight in lb. (x)	Frequency No. of boys (f)
60.5 — 70.5	1
70.5 — 80.5	5
80.5 — 90.5	9
90.5 — 100.5	14
100.5 — 110.5	15
110.5 — 120.5	4
120.5 — 130.5	2
	50

Mid values	f	$f. x_i$
65.5	1	65.5
75.5	5	377.5
85.5	9	769.5
95.5	14	1337.0
105.5	15	1582.5
115.5	4	462.0
125.5	2	251.0
	50	4845.0

$$\text{Mean} = \frac{\Sigma x.f}{\Sigma f} = \frac{4845}{50} = 96.9$$

$x - \bar{x}$	f	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
—31.4	1	985.96	985.96
—21.4	5	457.96	2289.80
—11.4	9	129.96	1169.64
— 1.4	14	1.96	27.44
8.6	15	73.96	1109.40
18.6	4	345.96	1383.84
28.6	2	817.96	1635.92
	50		8602.00

$$\text{Variance} = \text{Mean Square Deviation} = \frac{8602.00}{50}$$

$$= 172.04$$

$$\therefore \text{Root Mean Square Deviation} \quad \sigma = \sqrt{172.04}$$

$$= 13.1 \text{ Kg. per day.}$$

There are lot of difficulties in the calculation by the above method since it involves squaring of big numbers involving decimals and multiplication of square of such numbers by the frequencies. Therefore, we should think of alternative shortcut method.

B. Shortcut Method

There are two shortcut methods. In the case of first method, we use an assumed Mean say A. and find out the deviation of each value from the assumed Mean. Afterwards, we calculate the standard deviation of the deviation themselves, and the standard deviation of the deviation is same as the standard deviation of the original values.

Shortcut Method I

Let us assume 95.5 as A. The new table will be as follows:

Mid values x (1)	$d = x - 95.5$ (2)	f	$d.f$ (3)	d^2 (4)	$d^2 f$ (5)
65.5	-30	1	-30	900	900
75.5	-20	5	-100	400	2000
85.5	-10	9	-90	100	900
A - 95.5	0	14	0	0	0
105.5	10	15	150	100	1500
115.5	20	4	80	400	1600
125.5	30	2	60	900	1800
		Total	70		8700

Let us take 95.5 as the arbitrary value or A. Let a new variable d be calculated with the following formula.

$d = x - A = x - 95.5$ and these values are given in col. (2)

Let us calculate the Mean and Variance of 'd'.

$$\bar{d} = \frac{70}{50} = 1.4 \qquad V(d) = \frac{\Sigma d^2 f - \bar{d}^2}{n}$$

$$= 8700/50 - 1.4 \times 1.4$$

$$= 174 - 1.96$$

$$= 172{\cdot}04$$

$$\text{Standard deviation} = \sqrt{172{\cdot}04}$$

$$= 13{\cdot}1 \text{ kg. per unit.}$$

It can be argued in other way also. We have seen earlier that the Mean Square Deviation calculated from any arbitrary value will be greater than the variance calculated from the Mean and the difference will be equal to the square of the difference between the Mean and arbitrary value.

Mean Square deviation about A:

$$= V(x) + (\bar{x} - A)^2$$
$$= V(x) + (96.9 - 95.5)^2$$
$$= V(x) + (1.4)^2$$
$$= V(x) + 1.96$$
$$\therefore V(x) = 174 - 1.96 = 172.04$$

The above method of calculation with an assumed Mean *is called Changing the Base.*

Method II

In this process we change not only the base but also the scale. We adopt the following substitutions.

$$d = \frac{x - A}{C}$$

Where A is an arbitrary value and C is the class interval. In this process we *reduce each value to $1/C$ th of the original value* by dividing it by C. Let us calculate the standard deviation for the same continuous *frequency distribution.*

Mid values (in kg.) x_i	Frequency (f)
65.5	1
75.5	5
85.5	9
95.5	14
105.5	15
115.5	4
125.5	2

After subtraction of the arbitrary values, the residual values are reduced as follows. Let us assume 95.5 as the arbitrary value. The value of C is equal to the class interval namely 10.

$$\frac{65.5 - 95.5}{10} = \frac{-30}{10} = -3$$

$$\frac{75.5 - 95.5}{10} = \frac{-20}{10} = -2$$

$$\frac{85{\cdot}5 - 95{\cdot}5}{10} = \frac{-10}{10} = -1$$

$$\frac{95{\cdot}5 - 95{\cdot}5}{10} = \frac{0}{10} = 0$$

$$\frac{105.5 - 95.5}{10} = \frac{10}{10} = 1$$

$$\frac{115.5 - 95.5}{10} = \frac{20}{10} = 2$$

$$\frac{125.5 - 95.5}{10} = \frac{30}{10} = 3$$

The distribution of 'd' will be as follows and afterwards the same method that we have adopted for the computation of standard deviation of 'd' in the first method can be adopted here also. Thus we can find the standard deviation of 'd'.

x_i (1)	d (2)	f (3)	fd (4)	d^2 (5)	fd^2 (6)
65.5	-3	1	-3	9	9
75.5	-2	5	-10	4	20
85.5	-1	9	-9	1	9
95.5	0	14	0	0	0
105.5	1	15	15	1	15
115.5	2	4	8	4	16
125.5	3	2	6	9	18
	Total	50	7		87

NB : Col (5) can also be avoided. Col (6) can be directly computed by multiplying col. (2) and col. (4) also.

$N = \Sigma f = 50.$

$$\bar{d} = \frac{\Sigma f d}{N} = 7/50 = 0.14$$

$$\text{Variance of 'd'} = V(d) = \frac{\Sigma f d^2}{N} - \bar{d}^2$$

$$= \frac{87}{50} - 0.14 \times 0.14$$

$$= 1.74 - 0.0196$$

Standard Deviation (or)

$$\sigma^2 = \sqrt{1.7204}$$

$$= 1.31 \text{ Kg.}$$

We have already seen in the first method that the standard deviation of '*d*' is same as the standard deviation of '*x*' when there is only change of base. But in this process we have changed not only the base but also the scale by dividing it by *C* which is equal to 10 in this example.

Since each value of '*d*' is 1/10th of the corresponding of *x*. we can naturally expect that the standard deviation of '*d*' will also be equal to 1/10th of the standard deviation of '*x*'. In other words, the standard deviation of '*x*' will be 10 times (*C*) the standard deviation of '*d*'.

$$\text{Variance of } x = 100 \times \text{Variance of 'x'}.$$

$$V(x) = 100 \times 1\ 7204$$

$$= 172.04$$

Similarly standard deviation of x σ = 10 x Standard deviation of *d*.

$$= 10\ \sigma d$$

$$= 10 \times 1.31$$

$$= 13.1 \text{ Kg.}$$

In this method, the original values of 'x' namely, 65.5, 75.5, 85.5 etc. are reduced in size and they are replaced by —3, —2, —1, 0, 1, 2, 3 etc. which are small numbers whose squares can be written without referring to any tables. Because of this, lot of labour is saved in squaring and multiplication.

Formula: Standard Deviation of x = c × $S.D.$ of d.

$$\sigma x = c \times \sigma x$$

Method I: Let $d = x - A$

$$\therefore \bar{d} = \bar{x} - A$$

$$d - \bar{d} = (x - A) - (\bar{x} - A)$$

$$d - \bar{d} = x - A - \bar{x} + A$$

$$= x - \bar{x}$$

$$(d - \bar{d})^2 = (x - \bar{x})^2$$

$$\Sigma (d - \bar{d})^2 = \Sigma (x - \bar{x})^2$$

If we divide by N,

$$\therefore \frac{\Sigma (d - \bar{d})^2}{N} = \frac{\Sigma (x - \bar{x})^2}{N}$$

$$V(d) = V(x)$$

$$\therefore \sigma d = \sigma x$$

$\therefore$ Variance of 'd' = Variance of 'x'.

$\therefore$ Standard deviation of 'd' = Standard Deviation of 'x'.

Because of the change in the base, the value of Standard Deviation does not undergo any change.

Let us consider the change in the scale and in the base.

Proof: Let 'd' $= \dfrac{x - A}{C}$

$$\therefore \bar{d} = \frac{\bar{x} - A}{C}$$

$$\therefore \; d - \bar{d} = \left(\frac{x - A}{C}\right) - \left(\frac{\bar{x} - A}{C}\right)$$

$$= \frac{x - A - \bar{x} + A}{C} = \frac{(x - \bar{x})^2}{C}$$

$$\therefore \; (d - \bar{d})^2 = \left(\frac{x - \bar{x}}{C}\right)^2$$

$$\Sigma \, (d - \bar{d})^2 = \Sigma \left(\frac{x - \bar{x}}{C}\right)^2$$

If we divide by 'N' we get,

$$\frac{\Sigma \, (d - \bar{d})^2}{N} = \frac{\Sigma \, (x - \bar{x})^2}{N. \, C^2}$$

$$\text{Variance of '}d\text{'} = \frac{\text{Variance of '}x\text{'}}{C^2}$$

By the rule of cross-multiplication we get;

C^2 Variance of 'd' $=$ Variance of 'x'.

$$C^2 \, V(d) = V(x)$$

$$\therefore \; \sqrt{C^2 \, V(d)} = \sqrt{V(x)}$$

$$\therefore \; C \times \sigma \, (d) = \sigma \, (x)$$

RETROSPECT

As standard deviation occupies an important place in the study of statistics, it is better to have a retrospect of what we have studied. We have so far considered the computation of standard deviation for two types of data namely (1) Raw data or ungrouped data; (2) Grouped data or classified data or frequency distribution. In the case of frequency distribution also, we have considered two categories namely (1) Discrete frequency distribution; (2) Continuous frequency distribution. Thus three types of data: (1) Ungrouped data; (2) Discrete frequency distribution and (3) Continuous frequency distribution are considered.

Under each category, we have seen two methods (1) Direct method and (2) Shortcut method. In the case of continuous frequency distributions, two types of shortcut methods are adopted by using the substitution:

$$(i)\ \ d = x - A: \qquad (ii)\ \ d = \frac{x - A}{C}$$

Now let us see one example wherein, the advantage of each method and the accuracy of the value of the standard deviation calculated can be seen. For this purpose let us examine the weight of 50 bags or bundles of some grain. This example has been considered (1) in the case of classification of data; (2) in the case of calculation of Arithmetic Mean of continuous frequency distribution; (3) in the case of calculation of standard deviation of continuous frequency distribution.

The weights of the bundles in kg. are given below:

67, 75, 127, 80, 85, 83, 93, 97, 91, 98,

98, 94, 102, 100, 102, 104, 105, 105, 103, 102,

121, 114, 79, 72, 82, 87, 88, 98, 107, 103,

90, 92, 98, 118, 111, 110, 106, 97, 109, 108,

107, 76, 89, 85, 88, 97, 91, 98, 112, 106.

Method I

Let us consider this series as raw data and calculate the Standard Deviation by the shortcut method. ie: without calculating the deviation from the Mean. We shall calculate the Standard Deviation from the square of the original values.

The working is as follows:

x	x^2	x	x^2
67	4489	87	7569
75	5625	88	7744
127	16129	98	9604
80	6400	107	11449
85	7225	103	10609
83	6889	90	8100
93	8649	92	8464
97	9409	98	9604
91	8281	118	13924
98	9604	111	12321
98	9604	110	12100
94	8836	106	11236
102	10404	97	9409
100	10000	109	11881
102	10404	108	11664
104	10816	107	11449
105	11025	76	5776
105	11025	89	7921
103	10609	85	7225
102	10404	88	7744
121	14641	97	9409
114	12996	91	8281
79	6241	98	9604
72	5184	112	12544
82	6724	106	11236

$$\text{Total of } x = 4850$$

$$\text{Total of } x^2 = 478479.$$

$$\Sigma x = 4850 \quad \bar{x} = 97.$$

$$\Sigma x^2 = 478479 \; ; \; N = 50.$$

$$V = \frac{478479}{50} - 97 \times 97$$

$$= 9569{\cdot}68 - 9409 = 160.68$$

$$\sigma = \sqrt{160{\cdot}68} = 12.7 \text{ kg. per head.}$$

Method II

Let us construct the frequency distribution:

Class	Frequency
65·5 — 75·5	1
75·5 — 85·5	5
85·5 — 95·5	9
95·5 — 105·5	14
105·5 — 115·5	15
115·5 — 125·5	4
125·5 — 135·5	2
	50

The calculation of Standard Deviation from the frequency distribution was also given under continuous distribution. We have calculated the Standard Deviation as 12.7. Now, let us compare the result as follows:

	From ungrouped data	From grouped data
1. Mean	9·7 kg.	96 9 kg
2. Variance $= V$	160·68 kg	172·04 kg
3. Standard Deviation	12·7 kg. per head	13·1 kg. per head

Though there is not appreciable difference in the value of Mean calculated in both the cases, their differences in the value of the variance and the standard deviation are appreciable.

It is seen that the values of variance and Standard Deviation, calculated from the grouped data are greater than those calculated from the ungrouped data. Further, we also know that the measures calculated from the ungrouped data are correct since it does not involve any assumption. But in the case of grouped data, one assumption is involved, namely that all the members in a particular group or class are having weight or value equal to the mid-value of that class. Because of this assumption, we find the difference in the Mean calculated from the grouped data. Similarly we find differences in the value of V and σ calculated from the grouped data.

We know that the assumption is not correct and also that the difference in the value is due to the assumption which arises out of the classification of data. The number of classes also depends upon the class interval. Hence the difference is due to the class interval. Therefore, if we want to know the correct value of V or σ, we have to apply a correction factor based on the class interval.

Sheppard's correction

The correction to be applied is known as Sheppard's correction. It will be equal to $C^2/12$ where C is the class interval. Since the value of variance obtained from the grouped data will have an upward tendency, the correction factor has to be subtracted from the variance obtained from the grouped data to get the correct variance of the data.

$$C = 10.$$

$$\text{Correction factor } C^2/12 = 100/12 = 8{\cdot}33.$$

$$\text{Variance obtained } V(x) = \sigma^2 = 172{\cdot}04.$$

$$\therefore \text{ Corrected variance} = 172{\cdot}04 - 8{\cdot}33$$

$$= 163{\cdot}71$$

$$\textbf{Standard Deviation} = \sqrt{163{\cdot}71} = 12{\cdot}8$$

The correction factor has further reduced the difference.

		For ungrouped data	For grouped data	Corrected value
Variance	(V)	160·68	172·04	163·71
Standard Deviation	(σ)	12·7	13·7	12·8

The above method is more or less similar to the one considered earlier and the only difference is that we have considered the mid-values instead of all the raw data. But in this, considerabe calculation in squaring and multiplication is involved. This can be avoided if we take an arbitrary value and compute the variance.

Characteristics of the Frequency Distribution and the frequency Curve

Mean and Standard deviation are used as tools for comparing different distributions. Mean indicates the closeness of the values with one another while the standard deviation indicates the dispersion of the values from one another. For comparison of different distributions, it is not necessary that the total number of units in the different distributions should be equal. However, we can make the total number of units in each distribution equal by converting each frequency into percentage so that the total of all the frequencies will be equal to 100 in the case of all the distributions. By this method we can indirectly make the total frequencies of all the distributions equal. Further, the percentages of the frequencies of different classes can be expressed in terms of probabilities also so that the total of the probabilities of all the classes will be equal to 1 for all the distributions. In this respect also the total number of units in different distributions can be made equal.

We have already studied that we can draw frequency curve for frequency distribution. Therefore, different frequency distri-

butions can be compared by means of their frequency curves. The area of the curves will depend upon the total number of units in the distribution. As the total number of units differ from distribution to distribution the area of the curve will naturally differ from curve to curve.

However, if we consider the percentages of the frequencies or probabilities instead of either the actual frequencies or the actual number of units in different classes, we can make the area of the different curves equal since in all cases the total area will be equal to 100 or 1 since the total of all probabilities will be always equal to 1. This curve will be called **probability integral curve.**

We have shown that the different values will be distributed around the Mean. Some of them will be less than the Mean and some others will be greater than the Mean. However, the area of all probability curves will be equal since the total area represents 1.

But the shape of the curves may differ either in the height or in the width. Because the area is constant, as the height increases the width or the spread will decrease and vice versa. As the spread or dispersion decreases, the variation among the values of the different units decreases. If all the units have equal values, they will be equal to the Mean, then the curve will be a straight line erected at the value equal to Mean. The width of the curve can be divided into six divisions three on either side of the Mean. The values of portions on the left-hand side of this Mean will be less than the Mean and the portion on the right hand-side of the of the Mean will be greater than the Mean. In order to make the comparison more effective, each division on either side can be equal to 1. But the width of each division though considered as equal to 1 in all cases will differ in different curves depending on the value of the standard deviation of the concerned distribution. Generally numbers are started from *O*. Hence the value of the mean is taken as '*O*' and portions on left hand side are denoted by — 3, —2, —1 and in the right hand side are denoted by +1, +2, +3. This shows that the curves are distributed from —3 to +3. with '*O*' as the Mean in all cases. Normally the curves will be a well bell shaped one and such curves are called Normal Curves.

MOMENTS OF A FREQUENCY DISTRIBUTION

The main characteristics of a frequency distribution are expressed by certain constants called Moments which are calculated from the frequency distribution. Moments can be calculated from any arbitrary point. The moments calculated from the Arithmetic Mean are very important and they are called the **Central Moments.** We can also calculate the moment from the origin namely 'O' and they are called Raw Moment. Moments of any order can be calculated. The moments are denoted by the letter μ. The different orders of the moments are denoted by μ_1, μ_2, μ_3, μ_4 and so on. The rth Central Moment will be denoted by μ_r, and rth Raw moment will be denoted by μ'_r.

$$\text{rth Central moment } \mu_r = \frac{\Sigma (x_i - \bar{x})^r f_i}{N}$$

$$\text{rth Raw Moment } \mu'_r = \frac{\Sigma (x_i - O)^r f_i}{N}.$$

$$\text{First Central Moment } = \mu_1 = \frac{\Sigma (x_i - \bar{x}) f}{N} = O$$

$$\text{First Raw Moment } = \mu_1 = \frac{\Sigma x_i f_i}{N}$$

$$= \bar{x}$$

The Second Central Moment:

$$\mu_2 = \frac{\Sigma (x - \bar{x})^2 f_i}{N}$$

$$= V = \sigma^2 = \text{Variance.}$$

$$\mu^1_2 = \frac{\Sigma x_i^2 f_i}{N}$$

$$\mu_2 = \frac{\Sigma (x_i - \bar{x})^2 f_i}{N}$$

$$= \frac{\Sigma x_i^2 f_i}{N} - \bar{x}^2$$

$$= \mu'_2 - \mu'^2_1$$

$\therefore$ **Second Central Moment = Second Raw Moment — Square of the First Raw Moment.**

In this manner, the 3rd and 4th Central and Raw Moments can be easily calculated.

$$\mu_3 = \frac{\Sigma (x - \bar{x})^3 f_i}{N}$$

Let us consider the expansion of $(x - \bar{x})^3$

$$= x^3 - 3x^2\bar{x} + 3x\bar{x}^2 - \bar{x}^3$$

$$\Sigma (x - \bar{x})^3 = \Sigma x^3 - \Sigma 3x^2 \bar{x} + \Sigma 3x\bar{x}^2 - \Sigma \bar{x}^3$$

$$= \Sigma 3x^3 - 3\bar{x} \Sigma x^2 + 3\bar{x}^2 \Sigma x - \Sigma \bar{x}^3$$

$$= \Sigma x^3 - \Sigma 3x^2 \bar{x} + \Sigma 3\bar{x}^3 - \Sigma \bar{x}^3$$

$$= \Sigma x^3 - 3\Sigma x^2 \bar{x} + 2\Sigma \bar{x}^3$$

Dividing by "N" we get $= \mu'_3 - 3\mu'_2 \mu'_1 + 2\mu'^3_1$

It may be seen that the Central Moment is expressed in terms of the Raw Moment. In this manner the fourth Central Moment can also be written in terms of the Raw moment as follows:

$$\mu_r^4 = \frac{\Sigma (x - \bar{x})^4 f_i}{N}$$

$$\mu_4 = \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'^2_1 \mu'_2 - 3\mu'^4_1$$

SKEWNESS OF A FREQUENCY DISTRIBUTION

Mean and Standard deviation tell us about two important aspects of a frequency distribution. They are: (1) the central value and (2) the concentration of values around the central value. Another aspect of the frequency distribution is skewness.

A distribution can be classified as symmetrical or nonsymmetrical. It is based on the frequencies. A distribution is seem to be symmetrical when the frequencies are symmetrically or equally distributed at equal intervals, on either side of the Mode. In other words, frequencies at equal intervals on either side of the Mode are equal.

Let us examine this with the help of a distribution.

Value	Frequency
10	1
20	2
30	4
40	6
50	10
60	6
70	4
80	2
90	1

Mode of the distribution is 50.

The values 40 and 60 are at equal distance (10) from the Mode on its either side. Similarly, the values 30 and 70 are at equal distance or interval (20) on either side of Mode. The values 20 and 80 are at equal distance (30) on either side of the Mode. So also 10 and 90. We can form the following characteristics from the frequency column.

The values 40 and 60 which are at equal distance from the mode are having same frequency 6. In the same way the values 30 and 70 are having the same frequency 4. The values 20 and 80 are having the same frequency 2. The values 10 and 90 are having the same frequency 1. The values which are at equal distance on either side of the Mode are having equal or same frequency. This is called symmetrical distribution.

The curve of a symmetrical distribution

The curve of a symmetrical distribution will be as follows. It will be perfectly bell shaped curve.

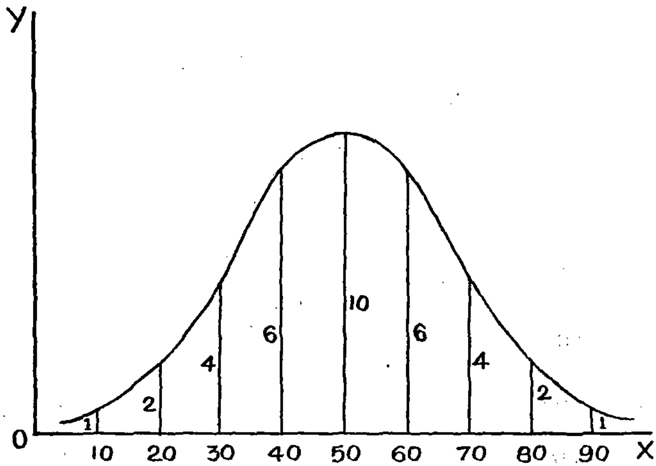


FIG. 16

Symmetrical Curve

Any curve of the above type is called symmetrical curve and the distribution of it is said to be symmetrical. Any departure from symmetry is known as skewness and the distribution is said to be skewed.

Properties of symmetrical distribution

1. In the case of symmetrical distribution the values of Mean and Median will coincide with the value of Mode. In other words, Mean, Median and Mode will be equal.

   Let a represent mean.

   M represent Median

   Z represent Mode.

2. Median will lie on the central point between the lower (Q_1) and upper (Q_3) Quartiles. The distance or

difference between M and Q_1 will be same as the distance or difference between Q_3 and M.

$$Q_3 - M = M - Q_1.$$

3. The sum of the positive deviations from Median will be equal to the sum of the negative deviations from Median.

4. The shape of the curve will be a perfect bell.

Skewness

The departure from the symmetry of a frequency *distribution* is known as skewness. For a perfectly symmetrical distribution all the odd central moments namely first, third, fifth etc. will be *equal to 'O'*.

For moderately symmetrical distribution the β_1 co-efficient will be a small positive quantity. For considerable departure from symmetry β_1 will be large. Hence β_1 *co-efficient can be taken* as a Measure of Skweness.

We know that $\beta_1 = \dfrac{\mu_3^{\,2}}{\mu_2^{\,3}}$

$$\therefore \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{\,3/2}}, \quad \beta_2 = \frac{\mu_4}{\mu_2^{\,2}}$$

The skewness may be either positive or negative depending upon the sign of μ_2.

Positive skewness

In a positive skewed distribution, the following characteristics will be noticed:

1. The number of items on the right side of the highest ordinate (height) of the curve will be more.

2. Median value will be greater than Mode.

3. Mean will be greater than Median. The ascending order of the values is (1) Mode, (2) Median, (3) Mean.

4. The frequency curve will have a long tail at the right side.

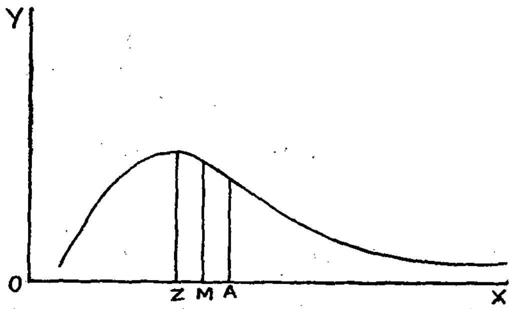


FIG. 17
Positive Skewness

Negative skewness

The following characteristics can be noticed in the case of negative skewness.

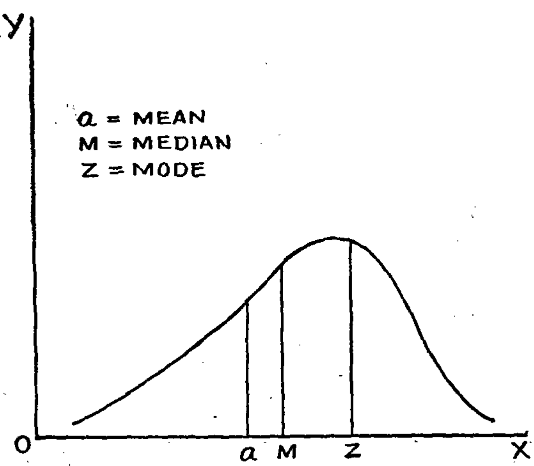


FIG. 18
Negative Skewness

1. The number of items on the left side of the highest ordinate (height) will be more.

2. Median will be greater than Mean.
3. Mode will be greater than Median. The ascending order of the values is i. Mean; ii. Median; iii. Mode.
4. The frequency curve will have a long tail on left.

Coefficient of skewness

In the case of symmetrical distribution Mean, Median and Mode will coincide. Therefore, in the case of skewed distribution they will not coincide. Hence their difference can be taken as a measure of Skewness.

Measures of skewness : Mean – Mode.

We have seen that Mean will be greater than Mode in the case of positive skewness. Therefore, the difference between the Mean and Mode will b, positive. In the case of Negative skewness, Mode will be greater than Mean. Hence, Mean—Mode will be negative.

Relative measure of skewness

In addition to β_1 co-efficient, a second measure of skewness is co-efficient of skewness denoted by C which is as follows:

$$C = \frac{\text{Mean} - \text{Mode}}{S.D.}$$

The above formula can be revised as follows:

$$C = \frac{3\ (\text{Mean} - \text{Median})}{S.D.}$$

This is known as Pearson's co-efficient of skewness. Since we have already seen that the difference between the Mean and Mod: will be equal to thrice the difference between Mean and Median, the second formula can be used. When there is perfect symmetry, Mean, Mode and Median will be equal to one another and consequently C will be equal to 'O'. If the mean is greater than the Median, or Mode, skewness sill be positive. If it is less than Median or Mode, the skewness will be negative.

3. A third co-efficient of skewness is given by the formula:

$$\frac{Q_3 + Q_1 - 2M}{2Q}$$

We know that $Q_3 - M = M - Q_1$

The difference of these will be expressed as a ratio to their sum.

$$\begin{aligned} \text{Difference} &= (Q_3 - M) - (M - Q_1) \\ &= Q_3 - M - M + Q_1 = Q_3 + Q_1 - 2M \\ \text{Sum} &= (Q_3 - M) + (M - Q_1) \\ &= Q_3 - Q_1 \\ &= 2Q. \end{aligned}$$

$$\therefore \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2M}{2Q}$$

This measure is known as Bowley's measure of Skewness. For a symmetrical distribution the distance of the Median from the Lower quartile and upper quartile will be equal $(M - Q_1)$ $(Q_3 - M)$. Therefore, the difference between these two quartiles from the Median divided by the Inter Quartile Range $(Q_3 - Q_1)$ is taken as a measure of skewness.

KURTOSIS

The flatness or peakedness of a frequency curve is knowu as Kurtosis. It depends upon the number of items near the Mode. We find the flatness or Peakedness of curve only with reference to Normal curve. Normal curve is an ideal symmetrical curve. Therefore, measure of Kurtosis will tell us how far a particular frequency curve is nearer to or away from the normal curve or how far the given frequency conforms to an ideal normal curve.

Among the various measures of Kurtosis the percentile measure of Kurtosis is the simplest. This is obtained by dividing the difference between the 90th percentiles and 10th percentiles by quartile deviation.

$$\text{Percentile Measure of Kurtosis} = \frac{\text{Difference between 90th and 10th percentiles}}{\text{Quartile Deviation.}}$$

Kurtosis is defined as $\frac{\mu_4}{\mu_2^{\,2}}$ ie $\frac{\mu_4}{\sigma^2}$ $B4$ and it is equal to β_2 $= \dfrac{P_{90} - P_{10}}{\dfrac{Q_3 - Q_1}{2}}$

For normal distribution $\beta_2 = 3 = \dfrac{2(P_{90} - P_{10})}{Q_3 - Q_1}$

If $\beta_2 > 3$ the peak is sharper and if $\beta_2 <_3$ the peak flatted.

For an ideal normal curve, the percentile measure of Kurtosis is equal to 3.8.

1. **Meso Kurtic:** Any distribution with percentile measure of Kurtosis equal to 3.8 the curve is called **Meso kurtic.**

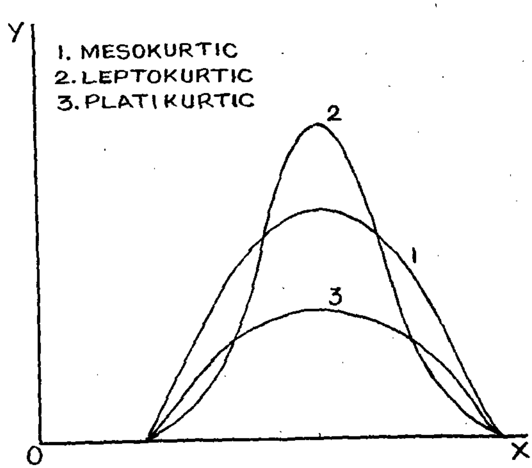


FIG. 19

Kurtosis

2. **Lepto Kurtic:** If the measure of percentile kurtosis of a curve is greater than 3.8, the curve is said to be Lepto Kurtic — has more peak at the tcp than the normal curve.

3. **Platy Kurtic:** If the percentile measure of Kurtosis of any distribution is less than 3.8, the frequency curve of the distribution is said to be platy kurtic or more flat at the top than the normal curve.

LORENZ CURVE

Lorenz curve is another method of studying the dispersion by means of graph. It is calculated from the cumulative frequency. In addition to the cumulative frequencies, we use the cumulative values also. The various steps involved are as follows:

(1) We should find the cumulative values. These values should be expressed in percentages.

(2) We should find the cumulative frequencies. The values should be expressed in percentages.

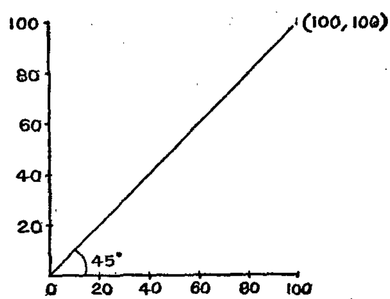


FIG. 20

Lorenz Curve

(3) The percentage values of the cumulative values should be plotted on the *X*-axis.

(4) The percentage values of the cumulative frequencies should be plotted on the Y-axis.

(5) In both the cases, the minimum value is 'O' and the maximum value is 100. If we join these points, the minimum and the maximum, will be a straight-line from the origin towards the opposite point. It will be a diagonal line.

Merits

It indicates how far the distribution has deviated from the average position. The average position will be indicated by the diagonal line joining the points with co-ordinates, probably (0,0) and (100,100).

Draw the Lorenz curve for the following distribution:

No. of workers	Total wages per mensem Rs.
3	2000
7	3000
12	4000
15	5000
13	6000
50	20000

No. of workers	Cumulative frequency	Percentage of C.F.
3	3	6
7	10	20
12	22	44
15	37	74
13	50	100

Wages (Rs)	Cumulative wages	Percentage of wages
2000	2000	10
3000	5000	25
4000	9000	45
5000	14000	70
6000	20000	100

The percentages of the cumulative values of frequency are as follows:

Percentages of cumulative workers	Percentages of cumulative wages.
6	10
20	25
44	45
77	70
100	100

The above figures can be plotted on the X and Y axes as follows:

X	:	6	20	44	77	100
Y	:	10	25	45	70	100

Plot the points and join them by a smooth curve. From the curve, we can also find out wages of above 25% or 50% of the workers. From them draw vertical line at points

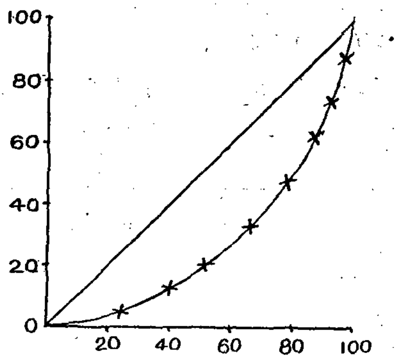


FIG. 21

Lorenz Curves

corresponding to 25% and 50% to cut the curve. The '*Y*' value of the point of intersection on the curve will give the percentage of of 25% of workers and 50% of the workers.

Exercise

(1) Calculate the Quartile deviation for the following data. (Marks)

52, 55, 57, 49, 54, 61, 64, 58, 63. 61.

(2) Calculate the semi inter quartile range.

	20	30	40	50	60	70
No. of students	5	7	20	8	4	6

(3) Calculate the Mean deviation for the following data.

132. 104, 166, 143, 175, 158, 179, 189, 125, 140.

(4) Calculate the Standard deviation and coefficient of variation for the following data.

(1) 28, 39, 45, 60, 65.

(2) 32, 53, 49, 28, 75.

(3) 280, 580, 350, 750, 625

(4) Calculate the *S.D.* and Coefficient of variation.

	x	f		x	f
(i)	25	3	(ii)	10	5
	35	4		20	8
	45	5		30	9
	55	8		40	7
	65	5		50	5
				60	6

(5) Calculate the *S.D* and Coefficient of variation.

	x	f		x	f
(i)	0 — 10	4	(ii)	0 — 25	11
	10 — 20	7		25 — 50	15
	20 — 30	8		50 — 75	28
	30 — 40	9		75 — 100	12
	40 — 50	5		100 — 125	14
	50 — 60	7			

CHAPTER III

CORRELATION

We have so far considered series having only one variable or one characteristic, for example, the weight of a person, or height of a person or the marks obtained by a student or the yield obtained from a plot etc. But in actual practice we may have to consider simultaneously more than one variable or characteristic at a time, for example, the height and weight of a person or quantity of fertiliser applied and the quantity of yield obtained. Sometimes each item may have three or more variables.

Correlation

The values of different variables may be inter-related. For example, the weight of a person may depend on the height of a person, or the height and weight of a person may depend upon the age of a person. The quantity of yield obtained may depend upon the quantity of fertiliser applied. The relationship between two or more variables is called the correlation and the variables are said to be correlated. Sometimes the relationship is also called as covariation.

Relationship

The term relationship can be used in three senses, namely (1) mutual relationship (2) cause and effect relationship and (3) general relationship.

Mutual relationship

The price and demand of a commodity have mutual relationship when the price of a commodity decreases, the demand of it may increase. Sometimes, when the demand for it increases, the price may also increase. Whenever there is a change in the

value of one variable there will be a change in the value of the other variable also. The changes in the values mutually depend upon the changes in the value of each other.

2. Cause and Effect

Whenever the quantity of fertiliser applied increases, the yield may also increase. In this case the cause for the increased effect in the production is the quantity of fertiliser applied. Increased rainfall may cause increase in production. In these cases the relationship is 'cause and effect'.

3. General Relation

The production of paddy may increase with the increase in the production of cotton. They may not have any direct relationship. However, the increase in rainfall may cause increase in the production of paddy, cotton and other agricultural commodities. In these cases, the production of one does not depend upon the other. However, they depend upon the increase in rainfall. Such relationship which has no inter-relationship, but however has a general relationship, with some other characteristics is called General Relationship.

But at present we shall consider only two characteristics for the sake of simplicity and easy understanding. Out of the two characteristics one will be considered as an independent variable while the other will be treated as a dependent variable. From the names given it may be understood that the value of the dependent variable depends upon the value of the independent variable. It may be clear that any variation in the value of the independent variable will also have an impact on the variation in the value of the dependent variable. If any such relationship exists in the changes of the values of the variables we can say that they are corretated. The independent variable will be denoted by the letter 'x' and the dependent variable by the letter 'y'.

Types of correlation

There are two types of correlation namely positive and negative correlation.

Positive correlation

Whenever the value of the independent variable increases, the value of the dependent variable of the corresponding unit will also increase or when there is a decrease in the value of the independent variable, the value of the dependent variable also decreases. In this case the changes in the values of the two variables are taking place in the same direction i.e. either both will increase or both will decrease simultaneoulsy. This is called positive correlation. Price and supply of a commodity can be best example of this type. Whenever the price of a commodity increases, the supply of the commodity will also increase or whenever the price of a commodity decreases the supply position will also decrease.

Negative correlation

On the other hand if the value of the dependent variable decreases when there is an increase in the value of the independent variable, or if the value of dependent variable increases when there is a decrease in the values of the independent variables, the changes in the values of variables are taking place in the opposite directions to one another and hence it is called Negative C rrelation. Price and demand of a commodity can be cited as an example. Whenever the price of the commodity increases, its demand may decrease and vice-versa.

Perfect positive correlation

The relationship between changes in the temperature and in the length of an iron bar can be a suitable illustration for perfect positive correlation. We know that the length of the iron bar will increase as the increase in the temperature and thus it indicates a positive correlation. But we also know that for every one degree rise in temperature, the length of the iron bar increases by a fixed length. This shows an uniform increase in the value of the dependent variable for a uniform increase in the values of the independent variable. We can otherwise say that the rates of changes in the values of both the variables are equal and this shows one to one correspondence in the rate of changes. Because of this one to one correspondence in the rate of changes

it can be said to be a perfect positive correlation and it can be indicated by +1. On the other hand if a similar but opposite rate of changes takes place in the value of the dependent variable corresponding to the rate of changes in the value of the independent value it will be called a perfect negative correlation and it can be indicated by—1. The volume of Gas at constant temperature decreases in a definite ratio when the pressure increases because '*PV*' is always constant. This shows that the indicator for correlation may vary between – 1 to +1 through '*O*'. If there is no correlation between the value of any two variables, we can say that indicator of the correlation is '*O*'.

Correlation can be studied by any one of the following methods.

1. Scatter Diagram; 2. Correlation graph; 3. Karl Pearson's Coefficient of Correlation.

The first two are graphical methods, and from these we can find out the nature of correlation. We can find out whether the correlation is positive or negative. i. e. Qualitative assessment of the correlation can be given but it cannot be expressed in quantitative measure from the graphical method. The coefficient of correlation gives a quantitative measure of the distribution.

Scatter Diagram and Correlation graph

Scatter Diagram

It is the simplest method of studying correlation between two relative variables. The type of correlation presents either positive or negative can be obtained with great ease from the diagram. The various values corresponding to each pair of x and y can be plotted in a xy-plane. We will find that the various points corresponding to each unit will be scattered through out the XY plane and this Diagram is called Scatter Diagram. Since this diagram depicts the values of two variates it is also called a bi-variate diagram. If the points tend to cluster themselves along well defined curves, the curves are called regression curves. In such cases an association between the 2 variable is suggested.

If the regression curves are straightlines we say the regression is linear.

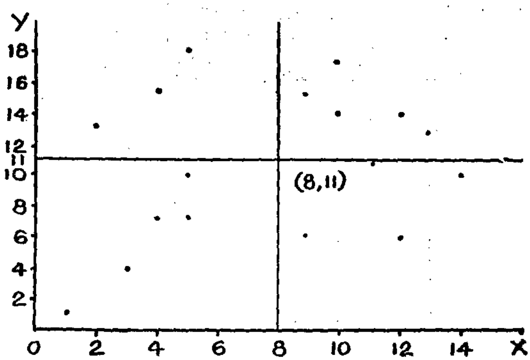


FIG. 22

Scattered Diagram

1. Positive Perfect Correlation

If all the plotted points or dots form a straight line running from left to right in the upward direction, the correlation is said

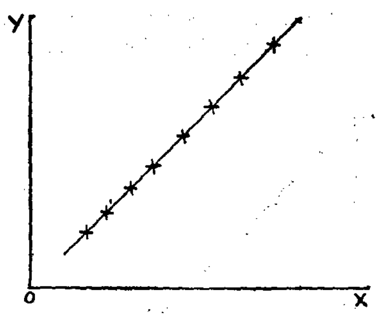


FIG. 23

Perfect Positive Correlation

to be **perfect positive.** The graph will be as given in the figure above:

2. Positive Correlation

If the points or dots are **scattered around** a straight line running from left to right in an upward direction, instead of all lying exactly on a straight line as explained before, the correlation is said to be positive. The figure will be as follows:

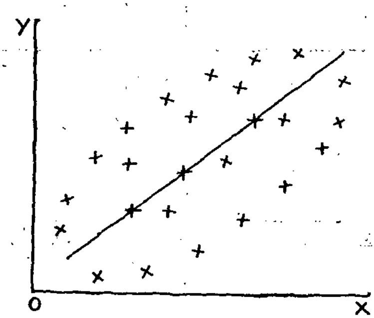


FIG. 24

Positive Correlation

3. Perfect negative correlation

If all the points or dots in a scatter diagram form a straight line running from right to left in a downward direction, the correlation is said to be perfect negative. The figure will be as follows.

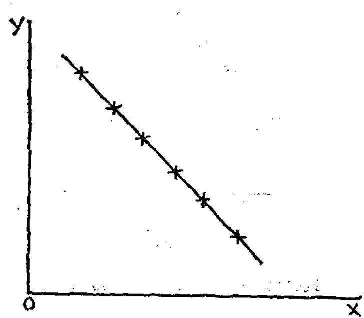


FIG. 25

Perfect Negative Correlation

4. Negative correlation

If the dots or points instead of all lying on a straight line as in the above figure, but scattered around a straight line, the correlation is said to be Negative. The figure will be as follows:

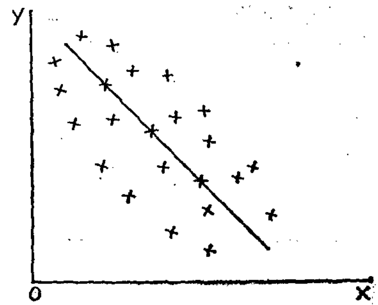


FIG. 26

Negative Correlation

5. No correlation

If the plotted points do not form a straight line but lie all over the plane as in the following figure, it will indicate the absence of any correlation.

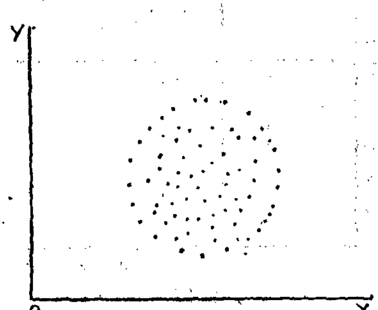


FIG. 27

No Correlation

Nature of correlation

The nature of correlation can also be determined from the scatter diagram by another method.

We can divide the total values of all *xs* and *ys* and find out their respective averages $\bar{x}$ and $\bar{y}$. We can also plot the point corresponding to the value $\bar{x}$ and $\bar{y}$ in the scatter diagram. By drawing two perpendicular lines, one to the x—axis and another to the y—axis, through the point corresponding to $\bar{x}$ and $\bar{y}$ and

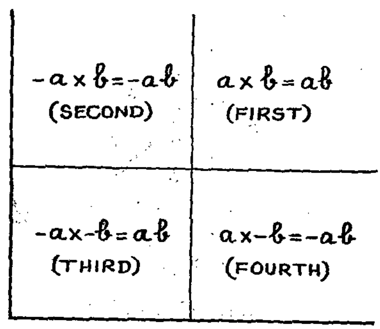


FIG. 28

extending the lines on either side of the point, we can divide the scatter diagram into four parts. Each part is called a quadrant. The quadrants are called as first, second, third and fourth quadrants.

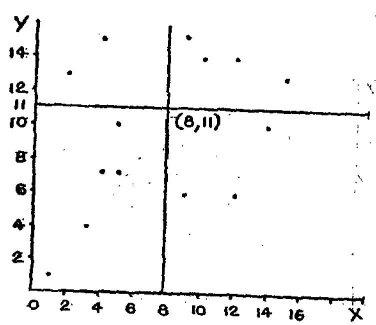


FIG. 29
Scattered Diagram

The number of points present in each quadrant should be counted. If the total number of points in the first and third

quadrants is greater than the total number of points in the second and fourth quadrants, it will indicate positive correlation. If the total number of points in the first and third quadrants is less than the total number of points in the second and fourth quadrants it will indicate negative correlation.

Correlation Graph

Correlation graph is used where variables are given with reference to a period of time. Time is marked on x axis and the values of the variables are marked on the y axis. Let us consider the following table and the graph drawn for it.

Year	Income Rs.	Expenditure Rs.
1970	100	90
1971	105	95
1972	115	100
1973	125	115
1974	150	120
1975	178	150
1976	190	175
1977	200	190
1978	205	200

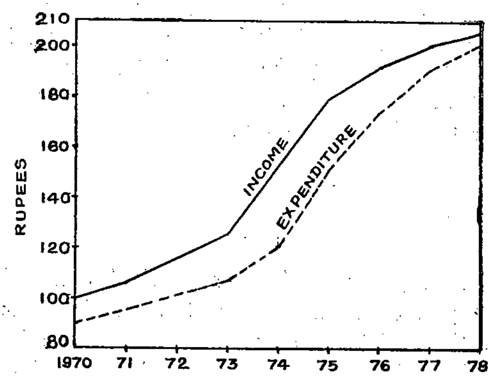


FIG. 30

Correlation Graph

A = Income curve; B = Expenditure curve.

The income and the expenditure are marked on Y-axis. All the points representing the income may be joined by means of a straight line. Similarly all the points representing the expenditure may be joined by means of a straight line which may be distinct from the straight line representing the income.

If the curves of the two variables are very close to each other, and if they move in the same direction, the variables are said to be positively related. On the other hand, if the curves of the two variables move in opposite directions, the variables are said to be negatively correlated. In the above example the curves for income and expenditure move in the same direction i.e. there is a rise in the height of one curve at a particular point of time corresponding to a rise in the other curve at that point of time, or there is a decline in one curve corresponding to a decline in the curve at a particular point of time. Hence the two variables are positively correlated. Correlation graph gives only an approximate idea of the correlation in the variables and it does not indicate the magnitude of the relationship.

A correlation graph can be drawn for the following data and then their relationship studied. This can be attempteid as an exercise by the students.

Year	Export	Import
		(Rs. in crores)
1970	100	80
1971	150	70
1972	175	65
1973	200	60
1974	250	50
1975	300	45

3. Coefficient of Correlation

Coefficient of correlation is calculated to study the extent or intensity or the degree of correlation exists between two variables. Correlation coefficient gives the degree of correlation in quantitative terms. Karl Pearson's coefficient of correlation is estimated by the following formula.

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N \sigma x \times \sigma y}$$

The above formula is indirectly based on certain assumptions given below:

i. The correlation between the two given variables is assumed to be linear.

ii. The forces affecting the two variables are assumed to be related to each other in a relationship of cause and effect.

iii. The various causes affecting the two variables are common to both.

In the above formula.

1. 'r' denotes the coefficient of correlation.
2. x and y are pair of values representing the two variables x and y.
3. $\bar{x}$ is the Mean value of the x variable.
4. $\bar{y}$ is the Mean value of the y variable.
5. σx is the standard deviation of the x values

$$\sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$$

6. $\sigma\, y$ is the standard deviation of the y values

$$\sqrt{\frac{\Sigma (y - \bar{y})^2}{n}}$$

7. 'n' is the number of items or the number of pairs. The above formula cn simplification may undergo the following changes and we can have three more types of formulae.

$$(1) \quad r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N\, \sigma x \times \sigma y}$$

In the above formula σx and σy can be substituted by their respective formulae. The formula will undergo changes as follows:

$$(2) \quad \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N \sqrt{\frac{\Sigma (x - \bar{x})^2}{N} \times \frac{\Sigma (y - \bar{y})^2}{N}}}$$

$$(3) \quad r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N \sqrt{\frac{\Sigma (x - \bar{x})^2\, \Sigma (y - \bar{y})^2}{N^2}}}$$

$$= \frac{\Sigma (x - \bar{x})\ (y - \bar{y})}{\frac{N \sqrt{\Sigma (x - \bar{x})^2\ (y - \bar{y})^2}}{N}}$$

(4) $$r = \frac{\Sigma (x - \bar{x}) \; (y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \; (y - \bar{y})^2}}$$

Generally $\Sigma (x - \bar{x}) \; (y - \bar{y})$ is called the product moment denoted by P_{11} and $\frac{\Sigma (x - \bar{x}) \; (y - \bar{y})}{N}$ is called the Mean Product Moment.

$$r = \frac{P_{11}}{\sigma x \cdot \sigma y}$$

Let us examine this with the help of one example. Calculate the coefficient of correlation for the following data.

x Weight in kg.	y Height in cms.
5	7
4	6
6	9
3	5
2	3
20	30

The various steps involved in the computation of 'r' are enumerated below:

Section I

i. Find the number of pair of items 'n' = 5.

ii. Find the Mean ($\bar{x}$) of the x values $= \frac{20}{5} = 4$ kg.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{20}{5} = 4$$

iii. Find the Mean ($\bar{y}$) of the values $\frac{30}{5} = 6$ kg.

$$\bar{y} = \frac{\Sigma y}{n} = \frac{30}{5} = 6$$

Section II

i. Find the deviation $(x-\bar{x})$ of each of the x—values from the Mean

ii. In the same way find the deviation $(y-\bar{y})$ of each of the Y — values from the Mean.

iii. The product of the deviation of 'x' from its mean and the deviation of 'y' from its mean may be calculated. The deviation of first value of 'x' may be multiplied by the deviation of the first value of 'y', and the second by the second value and so on.

$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
+1	1	1
0	0	0
2	3	6
−1	−1	1
−2	−3	6
0	0	14

$\therefore$ The product moment $\frac{(x-\bar{x})\ (y-\bar{y})}{n}$

$$= \frac{14}{5} = 2.8$$

Section III

1. Find the standard deviation of x'.

$$\sqrt{\frac{\Sigma (x - \bar{x})^2}{N}}$$

$(x - \bar{x})$	$(x - \bar{x})^2$
1	1
0	0
2	4
-1	1
-2	4
0	10

$$\sigma x = \sqrt{\frac{10}{5}} = \sqrt{2}$$

2. Find the standard deviation of y: $\sqrt{\frac{\Sigma (y - \bar{y})^2}{N}}$

$(y - \bar{y})$	$(y - \bar{y})^2$
1	1
0	0
3	9
-1	1
-3	9
0	20

$$\sigma y = \sqrt{\frac{20}{5}} = \sqrt{4} = 2$$

Product of $\sigma x. \sigma y = \sqrt{2} \times 2$

Section IV

Divide the Product Moment by the product of $\sigma x . \sigma y$

$$r = \frac{2.8}{\sqrt{2} \times 2} = \frac{2.8}{1.4 \times 2} = 1. \text{ (approximately)}$$

The various procedures can be summarised in the following table:

x (1)	y (2)	$(x-\bar{x})$ (3)	$(x-\bar{x})^2$ (4)	$(y-\bar{y})$ (5)	$(y-\bar{y})^2$ (6)	$(x-\bar{x})(y-\bar{y})$ (7)
5	7	1	1	1	1	1
4	6	0	0	0	0	0
6	9	2	4	3	9	6
3	5	—1	1	—1	1	1
2	3	—2	4	—3	9	6
Total 20	30	0	10	0	20	14
Mean 4	6	0	2	0	4	2.8

It may be seen that columns 3,4,5 and 6 are devised to calculate the standard deviation of x (σx) and standard deviation of y (σy). In this process we adopt the basic formula for the calculation of standard deviation.

$$(1) \quad \sigma x = \sqrt{\frac{\Sigma (x-\bar{x})^2}{N}} \qquad (2) \quad \sigma y = \sqrt{\frac{\Sigma (y-\bar{y})^2}{N}}$$

But we know that the standard deviation can be calculated by shortcut method by using the following formula. We can calculate the standard deviation from the original values themselves.

$$(1) \quad \sigma x = \sqrt{\frac{\Sigma x^2}{N} - \bar{x}^2} \qquad (2) \quad \sigma y = \sqrt{\frac{\Sigma y^2}{N} - \bar{y}^2}$$

For this we have to calculate the x^2 and y^2 columns and these two columns may replace columns 3, 4, 5 and 6.

The x^2 values can replace the columns containing $(x-\bar{x})$ and $(x-\bar{x})^2$

Similarly the y^2 values can replace the columns containing $(y-\bar{y})$ and $(y-\bar{y})^2$

The 7th column giving the value of $(x-\bar{x})\ (y-\bar{y})$ can be conveniently changed by another column giving the product value of x and y (xy) so as to facilitate the working. Let us see how this is possible.

7th column $= (x - \bar{x})\ (y - \bar{y})$

The total of col. (7) $= \Sigma\,(x - \bar{x})\ (y - \bar{y})$

Here $\bar{x}$ and $\bar{y}$ are constant;

$$= \Sigma\, xy - \bar{y}\Sigma\, x - \bar{x}\Sigma\, y + \Sigma\, \bar{x}\bar{y}$$

$$= \Sigma\, xy - \bar{y}.N\bar{x} - \bar{x}\, N\bar{y} + \Sigma\, \bar{x}\bar{y}$$

$$= \Sigma\, xy - N\,\bar{x}\,\bar{y} - N\,\bar{x}\,\bar{y} + N.\,\bar{x}\bar{y}$$

$$= \Sigma\, xy - N\,\bar{x}\bar{y}$$

$$\therefore\quad r = \frac{\Sigma\,(x - \bar{x})\,(y - \bar{y})}{N\sqrt{\dfrac{\Sigma\, x^2}{N} - \bar{x}^2}\ \sqrt{\dfrac{\Sigma\, y^2}{N} - \bar{y}^2}}$$

$$= \frac{\dfrac{\Sigma\, xy - N\,\bar{x}\,\bar{y}}{N}}{\dfrac{\sqrt{(\Sigma\, x^2 - N\,\bar{x}^2)\ (\Sigma\, y^2 - N\,\bar{y}^2)}}{N^2}}$$

$$\frac{\Sigma\, xy - N\,\bar{x}.\,\bar{y}}{N\sqrt{\dfrac{(\Sigma\, x^2 - N\,\bar{x}^2)\ (\Sigma\, y^2 - N\,\bar{y}^2)}{N^2}}}$$

$$= \frac{\Sigma\, xy - N\,\bar{x}\bar{y}}{\sqrt{(\Sigma\, x^2 - N\bar{x}^2)\ (\Sigma\, y^2 - N\bar{y}^2)}}$$

or $$\frac{\Sigma\, xy - N\,\bar{x}\,\bar{y}}{\sqrt{(\Sigma\, xx - N\,\bar{x}\,\bar{x})\ (\Sigma\, yy - N\,\bar{y}\,\bar{y})}}$$

The above expansion makes the position clear. Each factor of the denominator is identical with the numerator. The only difference is that in it first factor in the denominator we have substituted the values of x and $\bar{x}$ for y and $\bar{y}$ respectively occuring in the numerator.In the other factor the values of y and $\bar{y}$ are substituted for x and $\bar{x}$ respectively. If this point is kept in mind we can easily write the formula.

The above formula which is the most simplified contains only x, y, x^2, y^2 and xy columns. This does not involve calculation of $(x—\bar{x})$, $(y—\bar{y})$, $(x—\bar{x})^2$, $(y—\bar{y})^2$ a nd $(x—\bar{x})$ $(y—\bar{y})$. Let us adopt the new formula and calculate the correlation coefficient.

x	y	x^2	y^2	xy
5	7	25	49	35
4	6	16	36	24
6	9	36	81	54
3	5	9	25	15
2	3	4	9	6
20	30	90	200	134

$n = 5.$ $\bar{x} = \dfrac{\Sigma\, x}{n} = \dfrac{20}{5} = 4.$

$\bar{x}^2 = 4 \times 4 = 16.$

$\bar{y} = \dfrac{\Sigma\, y}{N} = \dfrac{30}{5} = 6$

$\bar{y}^2 = 6 \times 6 = 36.$

(1) $\Sigma\, x = 20.$ (2) $\Sigma\, y = 30.$ (3) $\Sigma\, x^2 = 90.$ (4) $\Sigma\, y^2 = 200.$
(5) $\Sigma\, xy = 134.$ (6) $N = 5.$

$$r = \frac{\Sigma xy - N\bar{x}\,\bar{y}}{\sqrt{(\Sigma x^2 - N\bar{x}^2)\ (\Sigma y^2 - N\bar{y}^2)}}$$

$$= \frac{134 - 5 \times 4 \times 6}{\sqrt{(90 - 5 \times 16)\ (200 - 5\times 36)}}$$

$$= \frac{134 - 120}{\sqrt{(90-80)\ (200-180)}} = \frac{14}{\sqrt{10 \times 20}}$$

$$= \frac{14}{10\sqrt{2}} = \frac{1.4}{1.4} = 1.$$

Calculate the value of '*r*' in the following case:

S.No. of the field	Quantity of fertilisers applied in the experimental plot kg. (x)	Quanitity of yield in the experimental plot kg. (y)
1	0	3
2	5	17
3	7	22
4	9	26
5	8	25
6	6	19
7	10	32
8	4	11
9	3	9
10	2	7

As we require the quantities of

Σx^2, Σy^2 and Σxy we have to compute the values of x^2, y^2 and xy as follows:

x	y	x^2	y^2	xy
0	3	0	9	0
5	17	25	289	85
7	22	49	484	154
9	26	81	676	234
8	25	64	625	200
6	19	36	361	114
10	32	100	1024	320
4	11	16	121	44
3	9	9	81	27
2	7	4	49	14
54	171	384	3719	1192

$N = 10.\quad \Sigma x = 54;\ \bar{x} = \dfrac{54}{10} = 5{\cdot}4\ \bar{x}^2 = \dfrac{5{\cdot}4}{10} \times \dfrac{5{\cdot}4}{10};\ \Sigma x^2 = 384$

$\Sigma y = 171 \quad \bar{y} = \dfrac{171}{10} = 17.1$

$\bar{y}^2 = \dfrac{17.1}{10} \times \dfrac{17.1}{10}\ \Sigma y^2 = 3719\ \Sigma xy = 1192$

$$r = \frac{\Sigma xy - N\bar{x}\bar{y}}{\sqrt{(\Sigma x^2 - N\bar{x}^2)\ \Sigma(y^2 - N\bar{y}^2)}}$$

Let us substitute the values of $\bar{x}$, $\bar{y}$, Σx^2, Σy^2 and Σxy in the expansion.

$$r = \frac{1192 - 10 \times 5.4 \times 17.1}{\sqrt{(384 - 10 \times 5.4 \times 5.4)\ (3719 - 10 \times 17.1 \times 17.1)}}$$

$= 0.998$

It shows that there is a high positive correlation between the application of fertiliser and the yield.

Shortcut Method

x	y
0	3
5	17
7	22
9	26
8	25
6	19
10	32
4	11
3	9
2	7

In the previous methods we had to find out the squares of each of the x and y values. Again we have to find the product of the values of x and y. This is very simple when x and y are small values. On the other hand if the values of variables are large, as in the case of y, squaring and multiplying will be very difficult. Hence we have to find out some shortcut method. When the values of x and y are reduced so that we can easily find out the squares and products without the help of mathematical tables calculating machines.

In this shortcut method we shift the base and reduce the values. Let us take 6 in the case of x values and 19 in the case of y-values

as arbitrary values. Let us change the x-value into 'u' and 'y' value into 'v' by the following substitution.

$$u = x - 6. \qquad v = y - 19.$$

Let us substitute the value of u, v, u^2, v^2, uv in the place of x, y, x^2, y^2, xy respectively in the formula for r.

$$r = \frac{\Sigma\, xy - N\, \bar{x}\, \bar{y}}{(\Sigma\, x^2 - N\, \bar{x}^2)\,(\Sigma\, y^2 - N\bar{y}^2)}$$

$$= \frac{\Sigma\, uv - N\, \bar{u}\bar{v}}{\sqrt{(\Sigma\, u^2 - N\bar{u}^2)\,(\Sigma y^2 - N\bar{y}^2)}}$$

u	v	u^2	v^2	uv
—6	—16	36	256	96
—1	—2	1	4	2
1	3	1	9	3
3	7	9	49	21
2	6	4	36	12
0	0	—	—	—
4	13	16	169	52
—2	—8	4	64	16
—3	—10	9	100	30
—4	—12	16	144	48
—6	19	96	831	280

$$\bar{u} = \frac{-6}{10} = -0{\cdot}6 \; ; \; \bar{v} = \frac{-19}{10} = -1{\cdot}9.$$

$$r = \frac{\Sigma\, uv - N\,\bar{u}\,\bar{v}}{\sqrt{(\Sigma\, u^2 - N\,\bar{u}^2)\;(\Sigma\, v^2 - N\,\bar{v}^2)}}$$

$$r = \frac{280 - 10 \times -0.6 \times -1.9}{\sqrt{(96 - 10 \times (-0.6)^2\;\;(831 - 10\,(-1.9)^2}}$$

$$= \frac{280 - 11.4}{\sqrt{(96 - 3.6)\;\;(831\text{-}36.1)}} = \frac{268.6}{\sqrt{92.4 \times 794.9}}$$

$$= \frac{268.6}{\sqrt{(92.4 \times 794.9}}$$

$$= \frac{268.6}{9.6 \times 28.2} = \frac{268.6}{270.7} = 0.992$$

It is seen that the conversion of values from x into 'u' and y into 'v' does not affect the value of the correlation ratio.

In the above shortcut method we have changed the base from 0 to 6 in the case of x and from 0 to 19 in the case of y values. The above conversion is in the following form namely $d = x—A$: We can also have other types of conversions where we can have a change in the scale instead of change in the base, or we can have change in the base as well as change in the scale as detailed below. In all these cases the value of the correlation co-efficient will not be affected.

(1) Change in the base : $d = x—A$

(2) Change in the scale: $d = \dfrac{x}{C}$.

(3) Change *in* the base and in the scale : $d = \dfrac{x - A}{C}$

Since we have two different values *xy* and it will be confusing if we use '*d*' for both. Hence we can use *u* and *v* for *x* and *y* respectively.

	x	y
1. Change in the base	$u = x - A$	$v = y - B$
2. Change in scale	$v = \frac{x}{c}$	$v = \frac{y}{d}$
3. Change in the base and scale	$u = \frac{x - A}{c}$	$v = \frac{y - B}{d}$

The students can try the methods 2 and 3 for any example.

Interpretation of Karl Pearson's correlation co-efficient

The value of Pearson's coefficient '*r*' always lies between —1 and +1.

When '*r*' is equal to +1, it indicates perfect positive correlation, when it is equal to —1, it indicates perfect negative correlation and when it is equal to 0, it indicates no correlation.

Merits

It reveals the nature of the correlation between two variables and at the same time it gives a numerical measure of the correlation.

Demerits

1. Whether the correlation between two given variables is linear or not, we assume it to be linear when we calculate the Pearson's Coefficient of correlation.

2. It involves much time to find out the correlation coefficient.

Caution about the correlation coefficient

It may be noted that correlation coefficient only expresse association and it does not itself tell anything about the causes

of the relationships of the variates. Because two variates are correlated, we cannot say whether the variation in one variate is the cause or the result of variation in the other variate. We cannot say that the association is due to the mutual dependence of two variates or due to common causes affecting both of them. Further a high value of correlation does not always indicate relationship of the two variates since such high values may be accidental also. In certain cases a high value of correlation may exist between two variates such as number of births in the hospital and yield of wheat and such correlations may be called spurious correlation or Nonsense correlation.

Correlation Table

For a single variable we have prepared a frequency table. Similarly for the simultaneous distribution of two variables we can prepare a table called correlation table. An illustration of correlation table showing the number of blocks distributed according to area and population is given below. Generally, the correlation table will be a two way classification.

Distribution of number of Blocks in Tamil Nadu according to area and population

(Population '000')

Area in sq. kilometre	Less than 40	40-60	60-80	80-100	Greater than100	Total
Less than 150	—	12	4	7	4	27
150 - 250	4	35	51	24	5	119
250 - 350	1	28	68	30	5	132
350 - 450	1	7	24	20	5	57
450 - 550	—	3	7	7	3	20
550 - 650	—	—	5	7	2	14
650 - 850	—	—	1	2	—	3
Greater than 850	—	—	—	—	2	2
	6	85	160	97	26	374

In a two-way table particulars of two variables will be recorded. If one of the variables or both the variables in a two-way table are qualitative, the table will be known as a contingency table. When both the variables are quantitative then the two-way table will be called correlation table.

Construction of a correlation table

From the raw data we can also prepare a correlation table. Let us prepare a correlation table for the following data.

Height (cms.)	Weight (kg.)
156	60
128	62
145	75
120	45
110	40
112	49
115	52
120	38
140	75
125	59
111	65
105	42
112	50
115	60
120	63
125	68
117	74
115	75
112	65
119	69
120	70
125	55
130	58
145	60
150	75

Let us find out the maximum and minimum values of x and y.

	x	y
Maximum value	156	75
Minimum value	105	38
	51	37

In the case of 'x' we can have 10 as the class interval and in the case of 'y' we can have 5 as the class interval. The various classes can be arranged as follows:

X/Y	35-40	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-80	Total
100-110		1								1
110-120		1	1	2		1	3	1	1	10
120-130	1		1		2	2	1	1		8
130-140					1					1
140-150						1			2	3
150-160						1			1	2
Total	1	2	2	2	3	5	4	2	4	25

Since there are 25 items, the grand total of the columns and the rows is equal to 25.

After the construction of a correlation table of the above type we have to classify the data. For this purpose we should

consider simultaneously both x and y values of each item. Let us take first the x value ie, 156 and it occupies the last class namely 150—160. But there are 9 columns under y. We have to consider the y-values. The Y-value is 60. It occpies col. 60-65 Hence we have to put a tally mark (/) in the square against 150-160 under x and 60-65 under y. Similarly we have to classify the data for all the items.

After transferring all the items by means of tally marks as in the case of classification of data, we have to count the number of tally marks in each square or block. The number of tally marks in each square will represent the frequency. The number of tally marks in each square is the frequency and they are noted in the square of the model table. Finally, a correlation table of the above type will emerge out from these data.

EXERCISE

1. Find out the correlation coefficient:

x	y
78	125
89	137
97	156
69	112
59	107
79	136

2. Find out the correlation coefficient:

x	y
50	80
70	90
30	50
10	40
90	190
120	30
80	70
30	90

3. Calculate the correlation coefficient:

(*i*)	x	y
	55	35
	65	45
	75	55
	35	25
	45	45
(*ii*)	28	23
	41	35
	40	33
	38	34
	35	31

(4) Calculate r for the following data.

(i) x	y	(ii) x	y
5	8	10	12
6	9	8	10
4	6	9	15
7	10	20	25
8	12	13	8

CHAPTER IV

REGRESSION

Regression is allied to correlation. It means the tendency to retain. In statistics, regression means the average relationship between the variables. The correlation gives the degree of relationship between the two variables and it is independent of the unit in which the original values are expressed. The square of the correlation ratio (r^2) gives the relative amount of variation in the dependent variable. But Regression describes the functional relationship between the two variables.

Purpose of Regression Analysis

It is clear that the value of one variable (generally dependent variable) can be estimated from the value of the other variable (independent) with the help of the functional relationship. After estimating the value of one variable with the help of the functional relationship we can also find out the deviation between the observed value and the estimated value and this can be otherwise called the error of our estimate. Generally this is similar to the standard deviation. Hence the error of estimate is called the standard error of estimate. Therefore, three questions are involved in this study.

1. To find out the degree of relationship between the two variables called Correlation (r).

2. To find out the functional relationship between the two variables called Regression.

3. To find out the difference between the observed value and the value computed with the help of the functional relationship and express it as a measure of standard error of estimates Sy^2

Regression can be expressed graphically or algebraically. Graphic representation of regression is called Regression line. The Algebraic representation is known as Regression Equation.

Regression Line

The average relationship between two variables is described by the regression lines. When the exact value of one variable is given, the most probable value of the other variable is shown by the regression line.

When there are two relative variables, there will be two regression lines, one for each variable with reference to the other. Suppose there are two variables x and y, there will be one regression line for x with reference to y. This line is known as the Regression line of x on y. Another line for y with reference to x will give the value of y for the corresponding value of x. This is known as Regression line of y on x.

Properties of Regression Lines

(1) If the correlation between the two given variables is perfectly positive, i.e. when the correlation coefficient 'r' is equal to $+ 1$, the two regression lines will coincide with each other. It means that there will be only one line instead of two lines. In such situation the regression line will be as follows:

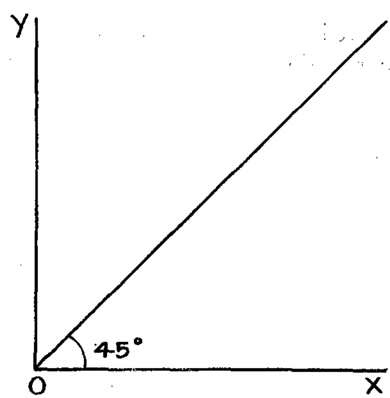


FIG. 31
Perfect Positive Regression

(2) On the other hand, if the correlation between two variables is perfectly negative, i.e: when the correlation co-efficient 'r' is equal to —1, the two regression lines will coincide and the line will be as follows:

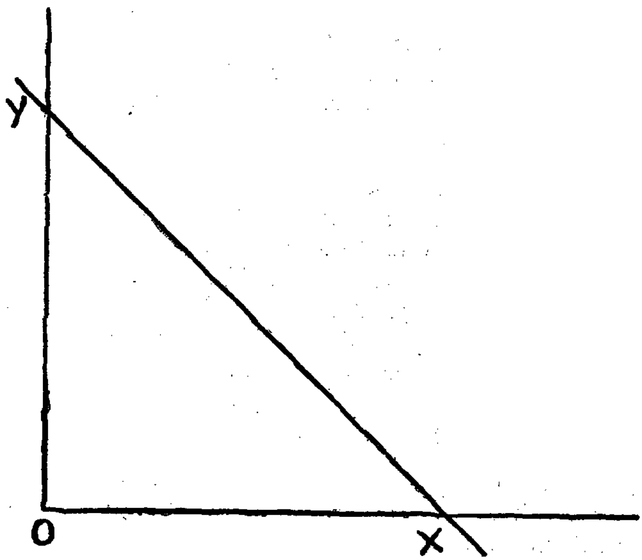


FIG. 32
Perfect Negative Regression

(3) When there is no correlation between the two variables, *i.e. when the correlation coefficient 'r' is equal to O,* the two

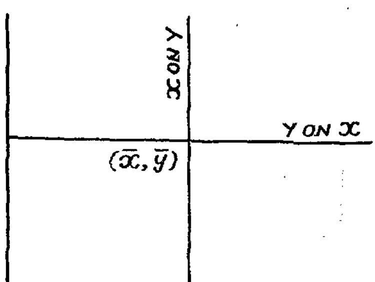


FIG. 33
No Correlation

regression lines will intersect each other at right angles. They will intersect each other at the point $(\bar{x}, \bar{y})$. The lines will be as in figure 33.

Regression Equation

Regression lines are generally given in algebraical expression known as Regression Equation. Regression equation will help us, to draw the Regression line. It will also give numerical method of finding out the best estimate of the value of one variable from the given value of the other variable.

When there are two variables x, y, then there will be two regression equations as follows:

1. Regression equation of x on y.
2. Regression equation of y on x.

The two regression equations will be as follows:

(1) Regression equation of x on y.

$$(x - \bar{x}) = \frac{r. \sigma x}{\sigma y} (y - \bar{y})$$

(2) Regression equation of y on x.

$$(y - \bar{y}) = \frac{r. \sigma y}{\sigma x} (x - \bar{x})$$

In these two equations

$\bar{x}$ = Arithmetic Mean of x.

$\bar{y}$ = Arithmetic Mean of y.

σx = the standard deviation of x.

σy = the standard deviation of y.

r = the correlation coefficient.

If the value of 'y' is given, we can use the equation (1) to find out the value of x corresponding to the value of y. If the value of x is given we can use the equation (2) to find out the value of y corresponding to the value of x.

Let us calculate the Regression Equation for the following data:

x :	0	5	7	9	8	6	10	4	3	2
y :	3	17	22	26	25	19	32	11	9	7

As we have to find out the value of the correlation coefficient 'r', we have to compute the value of x^2, y^2, and xy for each of the items and construct the columns for these values as we have done earlier when we studied about correlation ratio. The final table will be as follows:

x	y	x^2	y^2	xy
0	3	0	9	0
5	17	25	289	85
7	22	49	484	154
9	26	81	676	234
8	25	64	625	200
6	19	36	361	114
10	32	100	1024	320
4	11	16	121	44
3	9	9	81	27
2	7	4	49	14
54	171	384	3719	1192

$$\Sigma x = 54. \quad \bar{x} = \frac{54}{10} = 5.4$$

$$\sigma x = \sqrt{\frac{\Sigma x^2 - N\ \bar{x}^2}{N}}$$

$$= \sqrt{\frac{384}{10} - 5.4 \times 5.4}$$

$$= \sqrt{\frac{384 - 10 \times 5.4 \times 5.4}{10}} = \sqrt{\frac{384 - 291.6}{10}}$$

$$= \sqrt{\frac{92.4}{10}} = \sqrt{9.24} = 3.04$$

$$\sigma y = \sqrt{\frac{\Sigma\ y^2}{N} - \bar{y}^2}$$

$$\sigma y = \sqrt{\frac{3719}{10} - 17.1 \times 17.1}$$

$$= \sqrt{\frac{3719 - 2924.1}{10}} = \sqrt{\frac{794.9}{10}}$$

$$= \sqrt{79.49} = 8.9$$

$$r = \frac{\Sigma\ xy - N\ \bar{x}\bar{y}}{N\ \sigma x\ .\ \sigma y}$$

$$= \frac{1192 - 10 \times 5.4 \times 17.1}{10 \times 3.04 \times 8.91} = 0.998.$$

We have calculated the following values:

$x = 5.4$; $\sigma x = 3.04$; $\bar{y} = 17.1$ $\sigma y = 8.91$

$r = 0.992$

$\therefore$ The Regression Equation of X on Y is

$$(x - \bar{x}) = r. \quad \frac{\sigma x}{\sigma y}\ (y - \bar{y})$$

$$(x - 5.4) = 0.992 \times \frac{3.04}{8.94}\ (y - 17.1)$$

$$x - 5.4 = 0.34 \quad (y - 17.1)$$

$$x - 5.4 = 0.34y - 0.34 \times 17.1$$

$$x - 5.4 = 0.34y - 5.81$$

$$x = 0.34y - 5.81 + 5.40$$

$$= 0.34y - 0.41$$

$$x = 0.34y - 0.4$$

The Regression of y on x.

$$(y - \bar{y}) = r. \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$(y - 17.1) = 0.992 \times \frac{8.94}{3.04} \times (x - 5.4)$$

$$(y - 17.1) = 2.93 (x - 5.4)$$

$$y - 17.1 = 2.93x - 15.8$$

$$y = 2.93x - 15.8 + 17.1$$

$$y = 2.93x + 1.3$$

It is now clear that we can form the Regression Equation of y on x and the Regression equation of x on y provided we are given the following values:

$\bar{x}$, $\bar{y}$ σx, σy, r.

Let us write down the equation from the following data:

$\bar{x} = 5$; $\bar{y} = 7$; $\sigma x = 1$; $\sigma y = 2$; $r = 0.7$

(1) Regression Equation of x on y is:

$$(x - \bar{x}) = r. \frac{\sigma x}{\sigma y} (y - \bar{y})$$

$$(x - 5) = 0.7 \times \tfrac{1}{2} (y - 7)$$

$$(x - 5) = 0.35 \quad (y - 7)$$

$$x - 5 = 0.35y - 2.45$$

$$\therefore \; x = 0\,35y - 2.45 + 5$$

$$x = 0.35y + 2.55$$

(2) Regression of equation of y on x:

$$(y - \bar{y}) = r. \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$(y - 7) = 0.7 \times \tfrac{2}{1} (x - 5)$$

$$= 1.4 (x - 5)$$

$$= 1.4x - 7$$

$$y - 7 = 1.4x - 7$$

$$y = 1.4x.$$

Regression Coefficient

When we studied about the corrélation, we have come across the coefficient of correlation or the correlation coefficient which indicates the intensity of relationship of the two variables. Now when we consider the Regression we will have another coefficient called *Regression coefficient.*

While the coefficient of correlation indicates the intensity of the relationship, the Regression coefficient indicates the functional relationship of the two variables. The functional relationship is more important than the intensity of the relationship since the functional relationship helps us to calculate the value of one variable from the other. Hence the Regression coefficient plays greater part in the study of economic problems to estimate or forecast furure values of one item corresponding to the values of another item. Hence the students should clearly understand the subtle difference between the intensity of the relationship and the functional relationship.

We can quote one illustration from human relationship, even though it may not fully explain the difference.

When we say that *A* is a relative of *B*, we know that they are related. But we do not know to what extent the intensity of (closeness) their relationship exists. On the other hand, if we say that *A* is cousin of *B*, we understand that their relation is very close and not distant. This corresponds to the correlation coefficient.

We can further analyse the statement. Instead of saying that *A* is the cousin of *B*, we can say that *A* is the son of *B's* father's brother. This indicates their functional relationship. Perhaps this may correspond to the Regression coefficient. From this relationship we can give also the relationship of *A's* son and *B*'s son.

We have considered two equations

$$(x - \bar{x}) = r. \frac{\sigma x}{\sigma y} (y - \bar{y}) \;.......... (1)$$

$$(y - \bar{y}) = r. \frac{\sigma y}{\sigma x} (x - \bar{x}) \;.......... (2)$$

The first equation is the Regression equation of x on y and the second is the Regression equation of y on x.

Let us consider the factor

$r. \frac{\sigma x}{\sigma y}$ is the factor in the first equation. This is called the Regression coefficient of x on y. Similarly the factor $r. \frac{\sigma y}{\sigma x}$ in the second equation is the regression coefficient of y on x.

Regression coefficient of x on y: $r. \frac{\sigma x}{\sigma y}$

When there is one unit measurement change in the value of y, the value of x will be changed by the amount equal to $r. \frac{\sigma x}{\sigma y}$. In other words it indicates the amount of change in the value of x corresponding to one unit measurement change in the value of y.

Regression coefficient of Y on x $= r. \frac{\sigma y}{\sigma x}$:

When there is change of one unit measurement in the value of the x, value of y will be changed by the amount equal to $r. \frac{\sigma y}{\sigma x}$.

In other words $r. \frac{\sigma y}{\sigma x}$ will indicate the change in the value of y corresponding to one unit measurement change in the value of x.

We know that

$$r = \frac{\Sigma (x - \bar{x}) (y - \bar{y})}{\sigma x \ \sigma y}$$

$\therefore$ Regression co-efficient of x on y.

$$= r. \frac{\sigma x}{\sigma y} = \frac{\Sigma (x - \bar{x}) (y - \bar{y})}{\sigma x. \ \sigma y} \times \frac{\sigma x}{\sigma y}$$

$$= \frac{\Sigma (x - \bar{x}) (y - \bar{y})}{\sigma y^2}$$

Similarly the Regression co-efficient of y on x:

$$r. \frac{\sigma y}{\sigma x} = \frac{\Sigma (x - \bar{x}) (y - \bar{y})}{\sigma x^2}$$

Computation of correlation ratio (r) from Regression Coefficient

While we have two regression coefficients ie: (1) Regression coefficient of x on y and (2) regression coefficient of y on x, we have only one correlation ratio or correlation coefficient (r). Let us distinguish the two Regression co-efficients by the letters m_1 and m_2.

Let m_1 represent the regression coefficient of x on y.

$$m_1 = \frac{r. \ \sigma x}{\sigma y}$$

Let m_2 represent the regression coefficient of y on x.

$$m_2 = \frac{r. \ \sigma y}{\sigma x}$$

Let us now find out the value of their product.

$$m_1 \times m_2 = \frac{r.\ \sigma x}{\sigma y} \times r.\ \frac{\sigma y}{\sigma x} = r^2$$

$$\therefore\ \sqrt{m_1 \times m_2} = \pm\ r.$$

$$\sqrt{m_1\ \ m_2} = \pm\ r.$$

It is now clear that the correlation co-efficient is nothing but the square root of the product of the two regression co-efficients. Hence 'r' will have two values, one positive and another negative. If both the regression co-efficients are positive, take the positive root of '*r*'. When both regression co-efficients are negative, take the negative root of '*r*'.

Calculate the correlation ratio from the following details:

Regression co-efficient of x on y $= 0.9$

Regression co-efficient of y on x $= 0.4$

The product of the two regression co-efficients $= 0.9 \times 0.4$

$= 0.36$

$= r^2$

$\therefore$ Correlation ratio 'r' $= \sqrt{0.36} = 0.6$

From the above result, it is clear that we can calculate the average of the regression co-efficients provided we are given the value of '*r*' and the regression co-efficient of the variable.

Calculate the regression coefficient of y on x from the following data.

$r = 0.6$

Regression co-efficient of one variable on the second value $= 0.9$

Let the regression coefficient of the second variable on the first variable be equal to '*m*'.

$\therefore \quad m \times 0.9 = r^2$

$= 0.6 \times 0.6$

$= 0.36$

$\therefore \quad m = 0.36 \div 0.9$

$m = 0.4$

Regression co-efficient and Ratio of the Standard Deviation

In the previous case, we have multiplied the two regression co-efficients and proved that their product is equal to the square of the correlation ratio. r^2. Now let us consider the ratio of the two regression co-efficients:

$$\frac{\textbf{Regression co-efficient of } x \textbf{ on } y:}{\text{Regression co-efficient of } y \text{ on } x} \quad \frac{\left(r.\ \dfrac{\sigma x}{\sigma y} \right)}{r.\ \dfrac{\sigma y}{\sigma x}}$$

$$\text{Ratio } \tfrac{1}{2} = \frac{r.\ \sigma\ x}{\sigma\ y} \times \frac{\sigma\ x}{r.\sigma y}$$

$$= \frac{\sigma\ x^2}{\sigma\ y^2}$$

Regression equations and straight lines

We have studied in the previous chapter the construction of Regression Equations with the help of correlation co-efficient and the standard deviation of the two variables. The regression equations of

$$x \text{ on } y = (x - \bar{x}) = \text{r}.\ \frac{\sigma x}{\sigma y}\ (y - \bar{y})$$

$$y \text{ on } x = (y - \bar{y}) = r.\ \frac{\sigma y}{\sigma x}\ (x - \bar{x})$$

Afterwards we have considered the computation of the two regression co-efficients from the value of 'r' and the standard deviation of the two variables.

Regression co-efficient of x on y $=$ r. $\dfrac{\sigma x}{\sigma y}$

Regression co-efficient of y on x $=$ r. $\dfrac{\sigma y}{\sigma x}$

Generally it is usual to construct one Regression equation instead of two equations and one regression co-efficient instead of two regression co-efficients. The general convention is to construct the Regression co-efficient of y on x. This is due to the fact that the x values will be treated as an independant value and the y-value will be treated as a dependant value.

The methods we have considered earlier for the construction of Regression Equation may appear rather a more round about one. There is still another method in which the regression co-efficient will be first computed and afterwards the regression equation be formed. In this method, the important assumption made is that the relationship between the two variables is linear and consequently the Regression Equation will represent the equation to a straight line. Before we proceed further, let us study the equation to a straight line as we do in the co-ordinate geometry.

$y = mx + c$ is the general equation to a straight line where 'm' represents the slope of the angle made by the straight line with the X-axis, and 'c' represents the interception made by the straight line on the Y-axis from the origin.

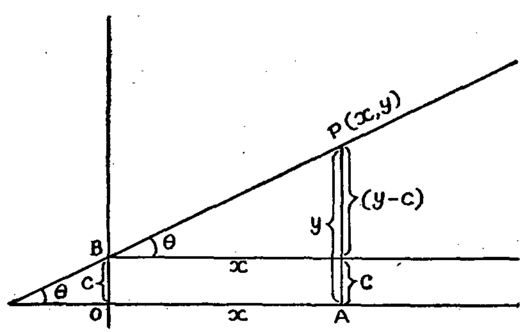


FIG. 34
Straight Line and Axes

After the assumption that the functional relationship between the two variables is linear i.e. in the form of an equation to a straight line, we should construct the exact equation giving the exact values for the two unknowns namely 'm' and 'c' in the equation $y = mx + c$. Once these two values are estimated the exact equation can be determined. We shall see how these two values are computed.

Since there are two unknowns namely m and c, we want two equations, preferably simultaneously, both involving m and c. Let us construct the two equations first.

Let $y = mx + c$. be the first equation (1)

Let us multiply throughout by x.

We get, $xy = mx^2 + cx$ and let this be the second equation (2)

$$y = mx + c \quad \ldots\ldots\ldots\ldots\ldots(1)$$

$$xy = mx^2 + cx \quad \ldots\ldots\ldots\ldots\ldots(2)$$

Let us now consider the first equation

$$y = mx + c.$$

There are many items involving x and y values and let us have five values in our example;

x: x_1, x_2, x_3, x_4, x_5.

y: y_1, y_2, y_3, y_4, y_5.

After the formation of the Regression equation $y = mx + c$ we can compute the y-value for each of the corresponding x values namely x_1, x_2, x_3, x_4, x_5 values for x in the original equation $y = mx + c$. For every x value, we have two y values, namely one value 'y' as given in the problem and another y-value as computed by us. Let us differentiate these two y-values as

y_o and y_c, where y_o means observed value of y and y_c means the computed value of y. The final position will be as follows:

x_i (1)	y_o (2)	Computed value (3)
x_1	y_1	$y_{c1} = mx_1 + C$
x_2	y_2	$y_{c2} = mx_2 + C$
x_3	y_3	$y_{c3} = mx_3 + C$
x_4	y_4	$y_{c4} = mx_4 + C$
x_5	y_5	$y_{c5} = mx_5 + C$

Let us add columns 2 and 3.

$$y_1 + y_2 + y_3 + y_4 + y_5 = mx_1 + mx_2 + mx_3 + mx_4 + mx_5 + C+C+C+C+C$$

$$= m(x_1 + x_2 + x_3 + x_4 + x_5) + 5Cs$$

$$\Sigma y = m \; \Sigma x + 5C \ldots\ldots\ldots\ldots\ldots (1)$$

Let us now multiply each of the 'y' values and computed values by the corresponding x values by x_1, x_2, x_3, x_4 and x_5 respectively. The equation will be as follows:

$$x_1 y_1 = mx_1^2 + Cx_1$$

$$x_2 y_2 = mx_2^2 + Cx_2$$

$$x_3 y_3 = mx_3^2 + Cx_3$$

$$x_4 y_4 = mx_4^2 + Cx_4$$

$$x_5 y_5 = mx_5^2 + Cx_5$$

Adding

$$x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5 = \Sigma\, xy$$

$$mx_1^2 + mx_2^2 + mx_3^2 + mx_4^2 + mx_5^2$$

$$+ Cx_1 + Cx_2 + Cx_3 + Cx_4 + Cx_5.$$

$$= m(x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2) +$$

$$+ C(x_1 + x_2 + x_3 + x_4 + x_5)$$

$$\Sigma\, xy = m\, \Sigma\, x^2 + C\, \Sigma\, x \quad \ldots\ldots\ldots(2)$$

Now we are having two equations as follows:

$$\Sigma\, y = m \quad \Sigma\, x + 5c \ldots\ldots\ldots\ldots\ldots (1)$$

$$\Sigma\, xy = m \quad \Sigma\, x^2 \quad + C \quad \Sigma\, x \ldots\ldots\ldots\ldots\ldots (2)$$

The number 5 appearing in first equation represents the number of items. Generally the number of items will be represented by N. Hence we can rewrite the equation as follows.

$$\Sigma\, y = m\, \Sigma\, x + Nc \ldots\ldots\ldots\ldots\ldots(1)$$

$$\Sigma\, xy = m\, \Sigma\, x^2 + C\, \Sigma\, x \ldots\ldots\ldots\ldots (2)$$

These two equations which help us to find out the values of m and c are called Normal Equations.

From the values given, we can easily have Σx, and Σy by adding their values. What we require further are Σx^2 and Σxy.

Of these, Σx^2 can be obtained by squaring each of the x values and adding them. Similarly Σxy can be obtained by multiplying each of the x values by the corresponding y-values and summing up the values. We can make the position clear with the help of an example.

S.No.	x	y	x^2	xy
1	0	3	0	0
2	5	17	25	85
3	7	22	49	154
4	9	26	81	234
5	8	25	64	200
6	6	19	36	114
7	10	32	100	320
8	4	11	16	44
9	3	9	9	27
10	2	7	4	14
Total	54	171	384	1192
	Σx	Σy	Σx^2	Σxy

Let us substitute these values in the above 2 normal equations.

$$54\,m + 10\,C = 171 \qquad (1)$$

$$384\,m + 54\,C = 1192 \qquad (2)$$

We can solve these two equations by means of simultaneous equations and get the values of m and c. In this process the co-efficient of any one of the variables namely m or c in both the equations can be made the same. Let us make the co-efficient of 'C' in both the equations same. For this purpose we can multiply the

first equation throughout by 54 and the second equation throughout by 10. These two equations will be changed as follows:

$$54 \times 54\,m + 54 \times 10\,C \quad = \quad 171 \times 54. \qquad (1)$$

$$384 \times 10\,m + 54 \times 10\,C \quad = \quad 1192 \times 10 \qquad (2)$$

In these two equations the quantity containing C are the same in both equations and having the same sign. Hence we can subtract equation (1) from (2) In this, the term containing C will vanish.

$$54 \times 54\,m = 171 \times 54 \quad (1)$$
$$384 \times 10\,m = 1192 \times 10 \quad (2)$$

$$2916\,m = 9234 \qquad (1)$$
$$3840\,m = 11920 \qquad (2)$$
$$2—1 \,/\, 924\,m = 2686$$
$$m = \frac{2686}{924} = 2.9074.$$

We can substitute this value for m in any one of the equations. Let us substitute the value in equation.

$$54\,m + 10\,C = 171$$
$$54 \times 2.9074 + 10C = 171$$
$$156.9996 + 10C = 171$$
$$10C = 171 — 156.9996$$
$$= 14.0004$$
$$C = \frac{14}{10} = 1.4$$

The required equation is $y = 2.9074\,x + 1.4$

We can summarise as follows:

(1) The line of best fit is called the Regression Line.

(2) The equation to the Regression Line is called Regression Equation.

(3) The constant '*m*', the coefficient of x is called Regression co-efficient

(4) The estimated value of y on the basis of the Regression Equation corresponding to a particular value of x is called Regression of y on x.

(5) The value of '*C*' is the value of y when x is equal to 0.

(6) The two equations used for computing the values of m and c are called the Normal Equations.

Computation of the dependent value with the help of the Regression Equation

After deciding about the functional equation we can estimate the value of '*y*' for each of the given values of x. In our problem the Regression Equation is

$$y = 2.9074 \quad x + 1.4$$

Let us consider this equation and find out the value of y for each of the values of x. For this purpose we have to substitute the value of x in the above equation. The values obtained are given below. Let the computed value be represented as Y_C

x	$Y_C = 2.9074\ x\ 1.4$
0	$Y_C = 2.9074 \times 0 + 1.4 = 1.40$
5	$= 2.9074 \times 5 + 1.4 = 15.94$
7	$= 2.9074 \times 7 + 1.4 = 21.75$
9	$= 2.9074 \times 9 + 1.4 = 27.57$
8	$= 2.9074 \times 8 + 1.4 = 24.66$
6	$= 2.9074 \times 6 + 1.4 = 18.85$
10	$= 2.9074 \times 10 + 1.4 = 30.47$
4	$= 2.9074 \times 4 + 1.4 = 13.03$
3	$= 2.9074 \times 3 + 1.4 = 10.12$
2	$= 2.9074 \times 2 + 1.4 = 7.21$
	171.00

We have two sets of 'y' value for each of the x values namely the observed value y_o. and the computed value y_c. We can test the validity of our equation indirectly by calculating the difference between the two sets of value for y.

y_o	y_c	$d = y_o - y_c$	$d^2 = (y_o - y_c)^2$
3	1.40	1.60	2.5600
17	15.94	1.06	1.1236
22	21.75	0.25	0.0625
26	27.57	—1.57	2.4649
25	24.66	0.34	0.1156
19	18.85	0.15	0.0225
32	30.47	1.53	2.3409
11	13.03	—2.03	4.1209
9	10.12	—1.12	1.2544
7	7.21	—0.21	0.0441
171	171.00	0.00	14.1094

We find that the total difference is 0.00 The total deviation is 0 because of the positive and negative deviations. This can be overcome by squaring the difference when all of them will become positive.

y_o	y_o^2	y_c	y_c^2	$(y_o - y_c)^2$
3	9	1.40	1.9600	2.5600
17	289	15.94	254.0836	1.1236
22	484	21.75	473.0625	0.0625
26	676	27.57	760.1049	2.4649
25	625	24.66	608.1156	0.1156
19	361	18.85	355.3225	0.0225
32	1024	30.47	928.4209	2.3409
11	121	13.03	169.7809	4.1209
9	81	10.12	102.4144	1.2544
7	49	7.21	51.9841	0.0441
171	3719	171.00	3705.2494	14.1094

We find that sum of the squares of the observed values of y is equal to the total of the sum of the squares of the computed values of y and the sum of the squares of the difference of the observed value and computed value.

$$\Sigma y_o^2 = \Sigma y_c^2 + \Sigma(y_c - y_c)^2$$

$$3719 = 3705.2494 + 14.1094$$
$$= 3719.35 \text{ or}$$
$$= 3719$$

The actual difference is only 0.35 which can be ignored for all practical purposes. Hence the difference can be taken as 0. Therefore the equation can also be rewritten as follows:

$$\Sigma(y_o - y_c)^2 = \Sigma y_o^2 - \Sigma y_c^2$$

Standard Error of Estimate

y_o - y_c is the deviation noticed between the observed value and estimated value of y.

$(y_o - y_c)^2$ = the square of the deviation between the two values.

$\Sigma(y_o - y_c)^2$ = the sum of the squares of the deviations between the observed value and the estimated value of all the items.

$\frac{\Sigma(y_o - y_c)^2}{N}$ = Average sum of or Mean squares of the deviation between the observed value and the estimated value of y.

$\sqrt{\frac{\Sigma(y_o - y_c)^2}{N}}$ = Average deviation between the observed value and the estimated value.

This can by writtin as follows;

$$V(d) = \sqrt{\frac{\Sigma d^2}{N} - \bar{d}^2} = \sqrt{\frac{\Sigma d^2}{N}} \text{ since } \bar{d} = 0. \text{ (almost)}$$

This is rather an average of the error noticed in the estimate. Hence it is called standard error of estimate. In short, this can be called the standard deviation of the deviation or differenc.

$$\sigma d = \sqrt{\frac{\Sigma d^2}{N} - \bar{d}^2} = \sqrt{\frac{\Sigma d^2}{N}} = \text{since } \bar{d} = 0$$

Importance of Regression Analysis

1. One of the important uses of regression is prediction or forecast. Hence greater importance is given to regression than correlation. It has great use in Economics.

2. Regression is used to estimate the study of supply and demand according to change in prices. It is also useful in estimating the likely increase in consumption and savings corresponding to increase in income.

3. It is used in the study of savings and investment.

4. Estimation of yield of crops due to the change in weather condition is being made with the help of regression.

5. Estimation of changes in public income due to changes in rates of taxation, estimation of changes in bank deposits and changes in bank loans due to changes in the rate of interest are being made by regression analysis.

Difference between correlation and Regression

1. Correlation gives the nature and degree of relation ship between two variables. But Regression gives the average change in the value of one variable corresponding to the change in the value of the other variable. Regression gives the exact or the functional relationship between the two.

2. The correlation between x and y is same as the correlation between y and x. But the Regression of x on y is not the same as the regression of y on x.

Exercise

(1) Fit a straight line for the following data.

x	y	x	y
5	8	10	12
6	9	8	10
4	6	9	15
7	10	20	25
8	12	13	8

(2) The Regression lines of Y on x and X on Y are given below. Find the correlation co-efficient between X & Y. Find also the ratio of $\sigma x : \sigma y$.

(a) $$Y = 0.80x + 25$$
$$X = 0.45y + 30$$

(b) $$X + 2Y = 5.$$
$$2X + 3Y = 8.$$

(3) From the given values find the 2 regression equations.

X	Y
70	120
80	130
90	150
60	110
50	100
70	130
60	120
60	100

(4) The height of fathers and son are given below. Find out the regression coefficient and estimate the height of the son when the height of the father is 164 cm.

Father's Ht.	**Son's Ht.**
160	180
165	160
162	170
158	180
168	160

(5) Find out the regression equations and obtain the best estimate of X when $Y = 7$ and the best estimate of Y when

$X = 5.$

$\bar{X} = 10,$ $\bar{Y} = 20$

$\sigma x = 1.5$ $\sigma y = 2$

$r = 0.7$

(6) Given the Mean of X and Y as 65 & 67.

Their standard deviation are 2.5 & 3.5.

The co-efficient of correlation is 0.8

(i) Find out the 2 regression equations

(ii) Find out the best estimate of

X when $Y = 70.$

CHAPTER V

RANK CORRELATION

The relationship between two variables can be studied in the following two ways.

1. We can study the relationship between the actual values of the two variables.

2. We can study the relationship between the ranks of the values of the two variables.

Ranking

An ordered arrangement of objects is called Ranking. Consider a set of individuals who are arranged in order according to some quality. The ranked data may arise from materials which are believed to be capable of measurements theoretically but at the same time they cannot be measured in practice. Intelligence, complexion etc., may come under this category.

Let us consider the marks obtained by 10 students in two subjects namely in language and science. Lef us also give ranks to them for each subject based on the marks given below.

a

If a student is uniformly good in both the subjects, he woul - get the same marks in both the subjects. If the same ideal situı ation is applicable to each and every student, each of them wil get the same marks in both the subjects and in other words each one would get same rank in both the subjects. Sometimes the marks obtained in both the papers may not be the same but differ from each other. But the ranks obtained by them in each subject may be the same. In such situations, the difference between the two ranks obtained by a single student in both the subjects will be 0. In the ideal situation explained above, the difference in

the ranks will be 0 in the case of each and every student and hence the total of the differences will also be 0.

S.No.	Marks obtained in language	Rank (R_1)	Marks obtained in Science	Rank (R_2)	Difference in the ranks ($R_1 - R_2$) (d)	Square of difference of the ranks. d^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1.	68	2	65	4	—2	4
2.	70	1	60	5	—4	16
3.	65	3	75	1	2	4
4.	55	5	50	7	—2	4
5.	50	6	70	3	3	9
6.	60	4	73	2	2	4
7.	48	7	55	6	1	1
8.	44	9	48	8	1	1
9.	40	10	45	9	1	1
10.	45	8	42	10	—2	4
TOTAL		55		55	0	48

But such an ideal situation may not exist. Naturally, there will be difference between the ranks obtained by a student. Even if the ranks obtained by each student in each subject is different, or even if there is difference in the ranks in the case of each and

every student, the total difference will be 0. Because of this peculiar situation a suitable formula has been designed for determining the rank correlation. The following formula is adopted where

$$r = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6\,\Sigma d^2}{n\,(n+1)\,(n-1)}$$

'*d*' represents the difference between the two ranks and *n* represents the number of students.

In the problem given

$$r = 1 - \frac{6 \times 48}{10 \times (100-1)} = 1 - \frac{6 \times 48}{10 \times 99}$$

$$= 1 - \frac{288}{990} = \frac{702}{990} = 0.71$$

The same illustration can be interpreted as the marks given by two examiners to the students on the same subjects.

Interpretation of Rank Correlation Co-efficient

1. Rank correlation co-efficient varies between -1 and $+1$
2. When R is equal to $+1$, there is complete agreement in the order of ranks in the case of both the variables.

Rank in first	**Rank in second**	d	d^2
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	0

$$R = 1 - \frac{6\ \Sigma d^2}{n\ (n^2 - 1)}$$

$$= 1 - \frac{6 \times 0}{5\ x\ (25-1)}$$

$$= 1 - 0$$

$$= 1.$$

3. When R is equal to — 1, there would be complete agreement in the reverse order of rank. But the ranks are in the opposite direction.

Rank	Rank	d	d^2
1	5	—4	16
2	4	—2	4
3	3	0	0
4	2	+2	4
5	1	+4	16
			40

$$R = 1 - \frac{6\ \Sigma d^2}{n\ (n^2 - 1)}$$

$$= 1 - \frac{6 \times 40}{5 \times 24} = 1 - 2 = \ -\mathbf{1}.$$

4. When R is equal to 0, there is no agreement at all in the order of ranks.

Exercise

(1) Calculate the Rank Correlation for the following data.

(i) Rank assigned		(ii) Marks assigned	
A	B	A	B
1	4	25	35
3	7	55	65
4	5	50	40
5	6	35	25
6	3	45	60
7	2	59	62
2	1	62	76
		28	42

CHAPTER VI

INDEX NUMBERS

Index number is a device for estimating the relative movement of values of statistical variables, in case where measurement of its actual movement is inconvenient or impossible. The measure of index number is appropriate when the variable in question is not stable in its composition. The basis of the method of index is relative.

The measures of central tendency give an average value of a group of figures, which gives an average of different items which are expressed in different units of measurements also. For example, the price of grain is expressed in terms of kilogram, the price of oil etc. in litres and the price of cloth in terms of metre. Still, we can give an average price of these items which are in different units of measurement, by means of index numbers.

Classification of Index Numbers

In the study of economics, various types of index numbers are used. Some of the important index numbers are as follows:

1. Price Index Number.
2. Quantity Index Number.
3. Value Index Number.
4. Index Numbers of Special Purposes.

Generally, we come across with the price index numbers and quantity index numbers. Therefore, let us study in detail the price index number, which is more commonly used in all spheres of economic activities.

Construction of Price Index Numbers

Suppose we are given the price of rice prevailing in two different periods of time, we can calculate the price index.

The price of rice in 1971 = Rs.2 per kg.

The price of rice in 1978 = Rs.3 per kg.

In such circumstances, the price prevailed in one period can be taken as the base for comparison and the price prevailed in the other period can be expressed as a ratio of the price in the base period. In our example, the price of rice in 1971 can be taken as the base year price while the price in 1978 may be taken as the price prevailed in the current year.

$$\text{The index number of price of rice in 1978} = \frac{\text{Price per unit measurement of rice in 1978}}{\text{Price per unit measurement of rice in 1971}} = \frac{\text{Rs. } 3}{\text{Rs. } 2} = \mathbf{1.5}$$

The price prevailed in the current year will be divided by the price prevailed in the base year and expressed in a ratio. In the above example, the ratio is 1.5. This means that the price of rice has increased by half of the price of the base year. In other words, the price has increased by 50%. Generally, the ratio will be converted into percentage by multiplying it by 100 so as to facilitate easy comprehension, comparison and also computation.

The formula for the price index is

$$\text{Price Index} = \frac{\text{Current year's price}}{\text{Base year's price}} \times 100$$

So in all our studies, the base year index is always taken as 100.

Example: Calculate the price index for the following:

Year	Price per unit of measurement Rs.	Price Index
1970	$4 = \frac{4}{4} \times 100 =$	100
1971	$5 = \frac{5}{4} \times 100 =$	125
1972	$6 = \frac{6}{4} \times 100 =$	150
1973	$7 = \frac{7}{4} \times 100 =$	175
1974	$8 = \frac{8}{4} \times 100 =$	200
1975	$10 = \frac{10}{4} \times 100 =$	250

Quantity Index

We have considered the price index. Let us now consider the Quantity Index. In the case of quantity index we will consider the quantity of the item in the base year and in the current year.

$$\text{Quantity Index} = \frac{\text{Current Year's Quantity}}{\text{Base Year's Quantity}} \times 100$$

Construct the quantity index in the following case taking 1972 quantity as the base.

Year	Quantity in quintals	Quantity Index
1972	12	$\frac{12}{12} \times 100 = 100$
1973	14	$\frac{14}{12} \times 100 = 116.7$
1974	15	$\frac{15}{12} \times 100 = 125.0$
1975	18	$\frac{18}{12} \times 100 = 150.0$
1976	20	$\frac{20}{12} \times 100 = 16.7$
1977	25	$\frac{25}{12} \times 100 = 208.3$

In our example we have considered only one item in the construction of Index Numbers. But in our day-to-day life, we come across with many items and naturally our index number should represent all the items. Therefore, we have to think of other methods which take into account of more than one item. There are different methods and we should examine each one separately.

Simple Aggregative Method

Suppose we are having five commodities and the price per unit of these commodities in different periods are given as follows, we can calculate the index number of the prices based on the Simple Aggregate Method.

Grain	Price per quintal (Rs.)		
	1960	1965	1970
Rice	51	60	65
Wheat	54	65	70
Cholam	48	60	70
Cumbu	45	65	70
Korra	42	50	55
TOTAL	240	300	330

Index Number of Prices in 1960 $= \frac{240}{240} \times 100 = 100 \cdot ^{\circ}$

Index Number of Prices in 1965 $= \frac{300}{240} \times 100 = 125.^{\circ}$

Index Number of Prices in 1970 $= \frac{330}{240} \times 100 = 137.5$

The total (aggregate) prices per unit of all the commodities in the current year is divided by the total (aggregate) prices per unit of all the commodities in the base year and the ratio is multiplied by 100.

$$I_n = \frac{\Sigma p_n}{\Sigma p_o} \times 100$$

In our example, we have taken different commodities. But the unit of measurement of all the commodities is same in all cases. Sometimes we may have commodities in differentunits. But the method of computation of index number is the same as before.

Commodities	Unit of measurement	Price per unit in	
		1960 Rs.	1965 Rs.
Rice	Kilogram	2.00	2.50
Oil	Litre	4.00	5.00
Cloth	Metre	5.00	7.25
Coconut	Number	1.00	1.25
		12.00	16.00

$$\text{Index Number} = \frac{16}{12} \times 100 = \mathbf{133.33}$$

The formula can be written as follows:

$$\frac{P_{1n} + P_{2n} + P_{3n} + P_{4n}}{P_{10} + P_{20} + P_{30} + P_{40}} \times 100$$

The numbers 1,2,3, 4 indicate the serial numbers of the commodities. n - indicates the current year; 0 - indicates the base year; P - indicates prices.

$$\frac{\Sigma P_n}{\Sigma P_o} \times 100$$

Disadvantage

The great disadvantage in this method is that all the items are given equal importance by taking only one unit under each item. But in actual life we give different importance to different items. Further, the prices of different commodities are expressed in different units. If we convert the prices of all the commodities for one uniform unit, say price per kilogram, the index number

worked out on the above basis will be completely different from the one calculated earlier. Therefore, the index number worked out on the basis of the si nple aggregate will not reflect the true and practical situation.

Simple Average of Relatives

We have already seen that the price index represents the relative changes in the prices. Hence we shall adopt the price relative of each commodity instead of the actual price of the different commodities.

We shall consider the same previous example:

Commodity	Unit	Price per unit in 1960 Rs.	Price per unit in 1965 Rs.	Price relative
1. Rice	Kilogram	2.00	2.50	$\frac{2.50}{2.00} = 1.25$
2. Oil	Litre	4.00	5.00	$\frac{5.00}{4.00} = 1.25$
3. Cloth	Metre	5.00	7.25	$\frac{7.25}{5.00} = 1.45$
4. Coconut	Number	1.00	1.25	$\frac{1.25}{1.00} = 1.25$
		12.00	16.00	5.20

The total cf price relatives $= 5.20$

Number of commodities $= 4$

The average price relativ.s $= \frac{5.20}{4} = 1.30$

Index No. $= 1.30 \times 100 = 130$.

Formula

$$\frac{\left(\frac{P_{n_1}}{P_{01}}+\frac{P_{n_2}}{P_{02}}+\frac{P_{n_3}}{P_{03}}+\frac{P_{n_4}}{P_{04}}\right)}{4}\times 100$$

$$\frac{\Sigma \frac{P_n}{P_o}}{N}\times 100 = \frac{\Sigma \frac{P_n}{P_o}\times 100}{N}$$

$N =$ indicates of commodities.

$n =$ indicates the current year.

o indicates the base year.

In the above method, we have calculated the index number by averaging the price relatives by arithmetic method. In other words, we have taken the Arithmetic Mean of the price relatives. Instead of the Arithmetic Mean, we can also adopt the Geometric Mean. In that case, the formula would undergo a change as follows:

$$I = \pi \left(\frac{P_n}{P_o}\right)^{1/N}\times 100$$

The price relatives calculated in the above previous example are given below:

Commodity	Price Relative
Rice	1.25
Oil	1.25
Cloth	1.45
Coconut	1.25

Geometric Mean of the price relatives:

$$I = \pi \left(\frac{P_n}{P_o}\right)^{1/N}$$

$$= (1.25 \times 1.25 \times 1.45 \times 1.25)^{\frac{1}{4}}$$

Log (G.M. of the price relatives) $= \log (1.25 \times 1.25 \times 1.45 \times 1.25)^{\frac{1}{4}}$

$$= \frac{\log 1.25 + \log 1.25 + \log 1.45 + 1.25}{4}$$

$$= 0.0969 + 0.0969 + 0.1614 + 0.0969$$

$$= \frac{0.4521}{4} = 0.1130$$

Geometric Mean $=$ antilog (0.1130)

$= 1.2970$

$\therefore$ Index No. $= 1.297 \times 100$

$=$ **129.7**

When we adopt the Arithmetic Mean, we get 130 as the index number. But when we adopt the Geometric Mean we get 129.7 as the index number. This is due to the fact that the Arithmetic Mean is always greater than the Geometric Mean.

Geometric Mean is preferred to the Arithmetic Mean

In the problem of Index Numbers, we are interested to know the relative changes rather than changes in the absolute values. Geometric Mean gives better result since it measures the relative change while the Arithmetic Mean measures the absolute change. Hence Geometric Mean is preferred to Arithmetic Mean in the calculation of Index Numbers. But the Arithmetic Mean is widely used for the simplicity of the computation.

Disadvantages

Out of the two disadvantages given in the case of simple aggregation method, the disadvantage due to different units for

different items is removed. However, the first defect, namely the uniform equal importance given to all the items, still continues. This defect also can be removed by giving due importance to the respective items. The method which gives the due respect to the individual item depending upon its importance is known as Weighted Method. The Index Number calculated by this method is known as Weighted Index Number.

Weighted Index Numbers

Index Numbers calculated by giving weights to various items according to their importance are called Weighted Index Numbers.

The weights to the various items can be given so as to bring out their economic importance. In certain cases, the quantities of the different items can be taken as the weight. In some other cases the values of the items can be taken as the weights. However, the following three kinds of weights are generally adopted.

TYPES OF WEIGHTS

1. Price Weights

In this case, the items included in the index numbers are given importance according to their prices.

2. Quantity Weights

The items included in the index numbers are given importance according to the quantity purchased or sold or consumed.

3. Value Weights

In this case the various items are given importance according to the expenditure incurred on those items.

There are two schools of thought in this process. Though there is no difference of opinion in the need for the adoption of weights, there is difference in approach in the adoption of weights. One school of thought prefers the current year prices or quantities or values as the case may be as the appropriate weights and another school of thought prefers the base year prices or quantities or values, as the case may be, as the appropriate weights. We should study these two approaches in detail.

Laspeyre's Index Method

In this method, the quantity of the base year is taken as the weight for calculating the Price Index Number.

$$I = \frac{\Sigma P_n q_o}{\Sigma P_o q_o} \times 100$$

For calculation of Index Number of Prices, we require (1) particulars of prices both in the current year and in the base year for the different commodities and (2) the quantities of the different commodities in the base year (q_o).

Aggregate Method

This method is known as Aggregate Method since $(p \times q)$ gives the values of items and consequently $\Sigma p \times q$ gives the total values.

Example: Calculate the Index Number of Prices from the following data:

TABLE NO. I.

Commodity	(Unit)	Quantity purchased in the base year	Price per unit in the base year	Price per unit in the current year
1	2	3	4	5
			Rs.	Rs.
Rice	kg	20	2.00	2.50
Oil	Litre	5	4.00	5.00
Cloth	Metre	10	5.00	7.25
Coconut	Number	12	1.00	1.25

From the above table we can construct the following table.

TABLE NO. II

Commodity	Total value in the base year	Total value in the current year
1	2	3
	Rs.	Rs.
Rice	20 × 2 = 40.00	20 × 2.50 = 50.00
Oil	5 × 4 = 20.00	5 × 5.00 = 25.00
Cloth	10 × 5 = 50.00	10 × 7.25 = 72.50
Coconut	12 × 1 = 12.00	12 × 1.25 = 15.00
	122.00	162.50

The prices of the commodities in the base year is multiplied by the quantities and the values are obtained as given in col. (2) of the Table Number II. Similarly, the values given in col. (3) of the Table No. II are obtained by the multiplication of prices in the current year by the base year quantities.

We get the following details from the Table Number II.

(1) $\Sigma P_n q_o$ = Rs. 162.50

(2) $\Sigma P_o q_o$ = Rs. 122.00

$$\therefore \text{ Index Number of Price} = \frac{162.50}{122.00} \times 100 = \mathbf{133.2}$$

Average of Ratios or Average of Price Relatives

Laspeyre's formula can be written as weighted average of price relatives also.

$$P_{on} = \frac{\sum \frac{P_n}{P_o} \times P_o q_o}{\sum \frac{P_o}{P_o} \times P_o q_o} \times 100$$

$$= \frac{\sum \frac{P_n}{P_o} \times P_o q_o}{\Sigma P_o q_o} \times 100$$

In this formula the base year's values of the commodities are taken as the weight. This method can be adopted if we are given the following details.

1. The base year values of the different commodities.
2. The base year prices of the commodities.
3. The current year prices of the commodities.

Commodities	Quantity	Base Year Price Rs.	Current Year Price Rs.
Rice	20 kg.	2.00	2.50
Oil	5 litres	4.00	5.00
Cloth	10 Metres	5.00	7.25
Coconut	12 Numbers	1.00	1.25

By multiplying the base year price and the base year quantity of each commodity, we can find the base year value of each commodity.

Rice	20×2	=	Rs. 40
Oil	5×4	=	Rs. 20
Cloth	10×5	=	Rs. 50
Coconut	12×1	=	Rs. 12
			Rs. 122

We shall construct the price relatives.

Commodity	Base Year Price	Current year price	Price relatives col. (3) ÷ col. (2)	Price relative × base year value.
(1)	(2)	(3)	(4)	(5)
	Rs.	Rs.		
Rice	2.00	2.50	1.25	$40 \times 1.25 = 50.00$
Oil	4.00	5.00	1.25	$20 \times 1.25 = 25.00$
Cloth	5.00	7.25	1.45	$50 \times 1.45 = 72.50$
Coconut	1.00	1.25	1.25	$12 \times 1.25 = 15.00$
			Total	162.50

(1) $\sum \frac{P_n}{P_o} \times p_o q_o = 162.50$

(2) $\Sigma p_o q_o = 122.00$

$$\text{Index Number} = \frac{162.50}{122.00} \times 100$$

$$= \mathbf{1.3.2}$$

We find now that the index numbers calculated with the help of the following two formulae are one and the same.

(1) $$\frac{\Sigma P_n q_o}{\Sigma p_o q_o} \times 100 = 133.2$$

(2) $$\frac{\sum \frac{P_n}{p_o} \times P_o q_o}{\Sigma P_o q_o} \times 100 = 133.2$$

This means both the formulae are one and the same.

Paasche's Index Number

In this method, the quantity of the commodity in the current year is used as the weight. We can calculate the Paasche's index number when we are given the following three details:

1. The base year price of each commodity.
2. The current year price of each commodity.
3. The current year quantity of each commodity.

$$\frac{\Sigma P_n q_n}{\Sigma P_o q_n} \times 100$$

Aggregate Method

In the above formula what we adopt is the total values of the commodities both in the current year and in the base year. But the quantity adopted is the current year quantity. Let us examine the same previous example with the quantity of the current year.

Commodity	Unit	Quantity	Price per unit Rs.	
			Base year	Current year
Rice	kg.	24	2.00	2.50
Oil	Litre	8	4.00	5.00
Cloth	Metre	12	5.00	7.25
Coconut	Number	16	1.00	1.25

From the above table we should construct another table giving the values of the commodities by multiplying the price by the quantity of the current year.

Commodity	Value in the base year Rs.	Value in the current year Rs.
Rice	$24 \times 2 = 48$	$24 \times 2.50 = 60$
Oil	$8 \times 4 = 32$	$8 \times 5 = 40$
Cloth	$12 \times 5 = 60$	$12 \times 7.25 = 87$
Coconut	$16 \times 1 = 16$	$16 \times 1.25 = 20$
	156	207

(1) Total value of the commodities in the current year $= \Sigma P_n q_n =$ Rs. 207

(2) Total value of the commodities in the base year $= \Sigma P_o q_n =$ Rs. 156

$$\therefore \text{Index Number} = \frac{207}{156} \times 100 = 132.7$$

Average of Ratios (or) Average Price Relatives

Paasche's formula can be written as a weighted average of price relatives.

$$\text{Index Number} = \frac{\sum \frac{P_n}{P_o} \times p_o q_n}{\sum \frac{P_o}{P_o} \times P_o q_n} \times 100$$

In this formula, the value of the current year quantity (q_n) at the base year price (p_o) is $P_o q_n$ is taken as the weight. This method can be adopted if we are given the following details.

(1) The base year price of the commodity.

(2) The current year price of the commodity.

(3) The value of the current year quantity of commodity at the base year price.

Commodity (1)	Quantity in the current year (2)	Base year price Rs. (3)	Current year price Rs. (4)
Rice	24 kg	2.00	2.50
Oil	8 litres	4.00	5.00
Cloth	12 metres	5.00	7.25
Coconut	16 nos.	1.00	1.25

By multiplying columns 2 and 3 we get the value.

Commodity	**Value** Rs.
Rice	48
Oil	32
Cloth	60
Coconut	16
	156

Commodity	Price relative	Price relatives × weight
Rice	1.25	$1.25 \times 48 = 60$
Oil	1.25	$1.25 \times 32 = 40$
Cloth	1.45	$1.45 \times 60 = 87$
Coconut	1.25	$1.25 \times 16 = 20$
		207

$$\text{Index Number} = \frac{207}{156} \times 100 = \mathbf{132.7}$$

We find that the index numbers calculated with the help of the following two formulae are one and the same.

(1) $$\frac{\Sigma\, P_n\, q_n}{\Sigma\, P_o\, q_n} \times 100$$

(2) $$\frac{\Sigma \dfrac{P_n}{P_o} \times p_o\, q_n}{\Sigma \dfrac{P_o}{P_o} \times p_o\, q_n} \times 100$$

Comparison of Laspeyre's Method and Paasche's Method

1. In both the formulae we use the prices prevailed in the base year and the current year.

2. The price relatives used are also same $\frac{P_n}{P_o}$.

3. In the case of Laspeyre's Method we have used base year quantity as the weight when we adopt the actual prices.

4. In the case of Paasche's Method, we adopt the curren year quantity as the weight when we adopt the actual prices.

5. When we adopt the price relatives, instead of the actual quantity we adopt the actual **expenditure incurred in the base year** as the weight in the case of Laspeyre's method.

6. When we adopt the price relatives instead of the actual quantity we adopt the **anticipated expenditure** in the current year and not the actual expenditure as the weight in the case of Paasche's Method. (Value of the commodity of the current year quantity at the base year price level) which is something imaginary and not actual.

Because of the basic differences of item 5 and 6, Laspeyre's method is preferable to the other.

Marshall - Egdeworth Index Method

A compromise of the Laspeyre's formula and Paasche's formula is also used to calculate the price index number. In this new method, sum of the quantities of both base year (q_o) and the current year (q_n) ie: ($q_o + q_n$) is used as the weight. The formula is given below:

$$\text{Price Index Number} = \frac{\Sigma P_n (q_o + q_n)}{\Sigma P_o (q_o + q_n)} \times 100$$

$$\text{(or)} \frac{\Sigma P_n q_o + \Sigma P_n q_n}{\Sigma P_o q_o + \Sigma P_o q_n} \times 100$$

In this method the weights (p_o and q_n) are changed every year and hence cannot be used for comparison. But in the case of Laspeyre's method the weights P_o q_o are fixed.

Hence the index numbers can be directly computed.

Commodity	Quantity		Prices in the	
	Base Year	Current Year	Base Year Rs.	Current Year Rs.
	q_o	q_n	P_o	P_n
			Rs.	Rs.
Rice (kg)	20	24	2.00	2.50
Oil (litre)	5	8	4.00	5.00
Cloth (metre)	10	12	5.00	7.25
Coconut (No.)	12	16	1.00	1.25

When we adopt the sum of the quantities of both the base year and the current year, the weight will be as follows:

Rice	20 + 24	= 44 kg.
Oil	5 + 8	= 13 litres
Cloth	10 + 12	= 22 Metres.
Coconut	12 + 16	= 28 Numbers.

After computing the weight, we can adopt the aggregate method.

Aggregate Method

Commodity	Quantity Weight	Value in the Base Year. Rs.	Value in the Current Year. Rs.
Rice	44	44 × 2 = 88	44 × 2.50 = 110.00
Oil	13	13 × 4 = 52	13 × 5 = 65.00
Cloth	22	22 × 5 = 110	22 × 7.25 = 159.50
Coconut	28	28 × 1 = 28	28 × 1.25 = 35.00
		278	369.50

$$\Sigma P_n (q_o + q_n) = 369.50$$

$$\Sigma P_o (q_o + q_n) = 278$$

$$\text{Index Number} = \frac{369.50}{278} \times 100$$

$$= 132.9$$

Quantity Index

We have so far considered the construction of Index Number of Prices. In all these cases, we have used the following items.

1. Price relative $\frac{P_n}{P_o}$

2. Any one of the following items as weight:

   i. Base year quantity (q_o)

   ii. Current year quantity (q_n)

   iii. Sum of the quantities of both base year and current year $(q_o + q_n)$

In the same manner we can also calculate the Quantity Index. In this we, use quantity relatives $\frac{q_n}{q_o}$ instead of price relatives. We use price as the weight.

The following items are used:

1. Quantity relative $\frac{q_n}{q_o}$

2. Any one of the following prices as weight.

   i. Base year price P_o.

   ii. Current year price P_n.

   iii. Sum of both base year and current year prices. (p_o+p_n)

The corresponding formula will be as follows:

1. $\frac{\Sigma q_n p_o}{\Sigma q_o p_o} \times 100$

2. $\frac{\Sigma q_n p_n}{\Sigma q_o p_n} \times 100$

3. $\frac{\Sigma q_n (p_n + p_o)}{\Sigma q_o (p_n + p_o)} \times 100$

Value Index

As we have calculated the Index Numbers for prices and quantities we can also calculate index number for values. When we multiply the price by the quantity we get the value.

$p_n q_n$: Value of the commodity in the current year.

$p_o q_o$: Value of the commodity in the base year.

$$\text{Index Number } V = \frac{\Sigma\, p_n q_n}{\Sigma\, p_o q_o} \times 100$$

Commodity	Quantity Base Year	Quantity Current Year	Prices Rs. Base Year	Prices Rs. Current Year
1	2	3	4	5
Rice	20	24	2.00	2.50
Oil	5	8	4.00	5.00
Cloth	10	12	5.00	7.25
Coconut	12	16	1.00	1.25

From the above table we can compute the values of the commodities in the base year and current year.

Commodity	Value in the Base Year Rs.	Value in the Current Year Rs.
Rice	40	60
Oil	20	40
Cloth	50	87
Coconut	12	20
	122	207

$$\text{Index Number} = \frac{207}{122} \times 100 = \mathbf{169.7}$$

Fisher's Ideal Index Number

A compromise of Laspeyre's formula and Paasche's formula is also adopted. This formula is known as Fisher's Ideal Index Number. It is the Geometric Mean of the above two formulae.

$$\text{Index Number} = \sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n}} \times 100$$

Ideal Index Number

Because of the following reasons it is called an ideal index number.

1. We know that the Geometric Mean is the best tool to indicate the relative changes. As index number is used to indicate the relative changes and Fisher's index is based on Geometric Mean, it is better than other index numbers.

2. The prices and quantities of both the base year and current year are considered in the construction of index number.

3. It satisfies the three tests namely, (i) The Commodity Reversal Test; (ii) The Time Reversal Test and (iii) The Factor Reversal Test.

However, this index number is not popular in use because of the laborious calculations involved in the construction.

Let us calculate the Fisher's Ideal Index Number for the following data:

Commodity	Quantities		Prices Rs.	
	Base year q_o	Current year q_n	Base year p_o	Current year p_n
Rice	20	24	2.00	2.50
Oil	5	8	4.00	5.00
Cloth	10	12	5.00	7.25
Coconut	12	16	1.00	1.25

1. We should first calculate the Index Number as per Laspeyre's Method.

$$I = \frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times 100$$

$$= \frac{(50 + 25 + 72.50 + 15)}{(40 + 20 + 50 + 12)} = \frac{162.50}{122.00} \times 100$$

$$= 133.2.$$

2. We should calculate the Index Number as per Paasche's Method:

$$I = \frac{\Sigma P_n q_n}{\Sigma p_o q_o} \times 100$$

$$= \frac{(60 + 40 + 87 + 20)}{(48 + 32 + 60 + 16)} \times 100$$

$$= \frac{207}{156} \times 100$$

$$= 132.7$$

3. We should calculate the product of these two index numbers:

$$133.2 \times 132.7$$

4. We should first find the square root of their product:

$$I = \sqrt{133.2 \times 132.7}$$

By taking log. on both sides we get,

$$\log I = \frac{\log 133.2 + \log 132.7}{2}$$

$$\log I = \frac{2.1245 + 2.1229}{2}$$

$$= \frac{4.2474}{2}$$

$$= 2.1237.$$

Index Number $=$ Antilog of 2.1237

$= 132.9$

Fixed Base and Chain Base Index Numbers:

Fixed Base

When we have values for a series of years we can take any one of the years as the base and calculate the index numbers for the values of the remaining years. This is known as Fixed Base Method. Index Numbers of this type can be compared more easily and effectively because of the common base. An example of this is given below:

Year	Price	Index Number
1960	20	100
1961	25	125
1962	40	200
1963	50	250
1964	70	350

We can take 1960 as the fixed base and work out the index number for the other years.

Chain Base Method

In this method, a common period is not taken as the base year. Instead, the year previous to the current year will be taken as the base for the succeeding year and because of this link this base is called chain base.

Let us calculate the index number with the chain base method:

Year	Value	Chain Base Index
1960	20	$\frac{20}{20} \times 100 = 100$
1961	25	$\frac{25}{20} \times 100 = 125$
1962	40	$\frac{40}{25} \times 100 = 160$
1963	50	$\frac{50}{40} \times 100 = 125$
1964	70	$\frac{70}{50} \times 100 = 140$

Conversion from Chain base to the Fixed Base

Chain base Index number can be converted into the Fixed Base Index Number.

Year	Chain Base Index	Fixed Base Index
1960	100	$\frac{100}{100} \times 100 = 100$
1961	125	$\frac{125}{100} \times 100 = 125$
1962	160	$\frac{125}{100} \times \frac{160}{100} \times 100 = 200$
1963	125	$\frac{125}{100} \times \frac{160}{100} \times \frac{125}{100} \times 100 = 250$
1964	140	$\frac{125}{100} \times \frac{160}{100} \times \frac{125}{100} \times \frac{140}{100} \times 100 = 350$

Merits of chain base method index number

1. People are generally interested in comparing the current year with the immediate proceeding year rather than with remote past. In such cases the chain base method is more useful.

2. It accommodates changes taken place in quick succession. It helps to add new items and also delete old items.

3. It helps to have a quick direct comparison of successive years.

4. It also helps to change the base as and when desired.

Tests of consistency for index number

When we studied the Fisher's Ideal Index Number, we have a reference to the following three tests

i. Commodity Reversal Test.

ii. Time Reversal Test.

iii. Factor Reversal Test.

It is said that index number which satisfies these three tests is an ideal index number. Let us see in detail about each of the tests.

1. Commodity Reversal Test

It has been said earlier that in the construction of index numbers we should consider not one item but many items. These items may be arranged in some order. If the order of arrangement of these items is changed, and the index number is worked out for the revised order, there will not be any difference between the original index number and the revised index number. This is due to the fact that there is nochange in the item either by addition or deletion. But the only change effected is in the order in which the items are considered. The change in order will not have any effect in value of the index number. Therefore, some are of the opinion that this will not constitute a test at all.

2. Time Reversal Test

We have seen that there are two periods, namely current year and base year, in the construction of index number. Generally the index number for the current year will be calculated with reference to the base year. Similarly we can also calculate the index number for the base year with reference to the current year. What we normally expect is that the one will be the reciprocal of the other. In other words, the product of these sets of index numbers will be 1.

The above condition will be satisfied if only one item is alone considered for the construction of index number. If more than one item is considered and also if the arithmetic average of the relatives are considered, this time reversal test will not be satisfied in the case of the two above formulae. This may be obvious and this can be verified with the help of the formula itself.

(a) Base year quantity as weight (Laspeyre's Method)

$$I = \frac{\Sigma P_n q_o}{\Sigma p_o q_o}$$

Let us interchange n and o in the formula.

The formula will be revised as

$$I' = \frac{\Sigma P_o q_n}{\Sigma p_n q_n}$$

$$I \times I' = \frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_o q_n}{\Sigma p_n q_n}$$

which is not equal to 1.

(b) Current year quantity as weight (Pasches' Method)

$$I = \frac{\Sigma p_n q_n}{\Sigma p_o q_n}$$

Interchange the letters n and o.

$$I' = \frac{\Sigma P_o q_o}{\Sigma P_n q_o}$$

$$I \times I' = \frac{\Sigma p_n q_n}{\Sigma p_o q_n} \times \frac{\Sigma p_o q_o}{\Sigma p_n q_o}$$

which is not equal to 1.

(c) **Fisher's Index Number**

$$\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p^n q_n}{\Sigma p_o q_n}} \qquad \dots\dots\dots\dots \quad (1)$$

Let us interchange the two pericds n and o and verify the formula with reference to Time Reversal test.

The formula will be

$$\sqrt{\frac{\Sigma p_o q_n}{\Sigma p_n q_n} \times \frac{\Sigma p_o q_o}{\Sigma p_n q_o}} \qquad \dots\dots\dots\dots \quad (2)$$

Multiplying (1) and (2)

$$\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n} \times \frac{\Sigma p_o q_n}{\Sigma p_n q_n} \times \frac{\Sigma p_o q_o}{\Sigma p_n q_o}}$$

Cancelling the common terms both in the numerator and denominator we get 1. The square root of 1 is also 1.

3. Factor Reversal Test

In the construction of Price Index numbers we are using two factors namely Price (p) and quantity (q). In the case of price Index Numbers, quantity is used as weight.

If we interchange the two factors, i.e: if we replace P by q and 'q' by P, we will get another set of index numbers which will be the index number of quantity with price as the weight.

If we multiply these two index numbers, i.e. the index number of Price and the index number of quantity the product should represent index number of the total expenditure, since the product of price and quantity will give the total value of the commodity. Perhaps this condition will be satisfied when we consider only one item. But if more items are considered as in the case of construction of index numbers and that too where Arithmatic average is adopted, this condition will not be satisfied. This can be verified. We can examine the two different formulae used for the construction of index number.

(a) With base year quantity as the weight (Laspeyre's Method)

$$IP = \frac{\Sigma p_n q_o}{\Sigma p_o q_o}$$

Let us replace P by q and q by P and the formula will become

$$IQ = \frac{\Sigma q_n P_o}{\Sigma q_o p_o}$$ which will be the index number of quantity

with base year prices as weight. The product (P) of the two index numbers.

$$p = IP \times IQ = \frac{\Sigma\ p_n q_o}{\Sigma\ p_o q_o} \times \frac{\Sigma\ q_n p_o}{\Sigma\ p_o q_o}$$

The product is not equal to 1.

(b) With current year quantity as weight (Paasche's Method)

$$IP = \frac{\Sigma\ p_n q_n}{\Sigma\ p_o q_n}$$

Interchange the fractiors p and q. The formula will become

$$IQ = \frac{\Sigma\ q_n p_n}{\Sigma\ q_o p_n}$$

This will be the index number of quantity with the Price of the current year as weight.

The product $IP \times IQ$

$$\frac{\Sigma\ p_n q_n}{\Sigma\ p_o q_n} \times \frac{\Sigma\ q_n p_n}{\Sigma\ q_0 p_n}$$

and this product will not be equal to the index number of the expenditure which will be denoted by the formula

$$\frac{\Sigma\ P_n q_n}{\Sigma\ P_o q_o} = P$$

The product is not equal to 1.

We have so far seen that the present two formulae adopted for the construction of index numbers are not satisfying the time reversal and factor reversal tests. But it can be proved with the help of the formula that the Fisher's index numbers will satisfy these two conditicns and perhaps this may be the reason that this index number is called an ideal index number.

Let us test this formula with reference to factcr reversal test.

$$\text{Formula} = \sqrt{\frac{\Sigma\ p_n q_o}{\Sigma\ p_o q_o} \times \frac{\Sigma\ P_n q_n}{\Sigma\ P^o q_n}}$$

Inter change the factors p and q. The formula will become

$$\frac{\Sigma q_n p_o}{\Sigma q_o p_o} \times \frac{\Sigma q_n p_n}{\Sigma q_o p_n}$$

product will become

$$\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n} \times \frac{\Sigma q_n p_o}{\Sigma q_o p_o} \times \frac{\Sigma q_n p_n}{\Sigma q_o p_n}}$$

$$\sqrt{\frac{\Sigma p_n q_n}{\Sigma p_o q_o} \times \frac{\Sigma q_n p_n}{\Sigma q_o p_o}}$$

$$\frac{\Sigma p_n q_n}{\Sigma p_o q_o}$$ which is nothing but the index number for the expenditure.

At present, index numbers are being published from various Departments. Mention may be made about the cost of living index numbers, consumer price index numbers, wholesale price index numbers, Index numbers for Agricultural Production, Index numbers for Industrial production etc.

Construction of Cost of Living Index Number

We know that the prices of various commodities are not constant and they are going on changing from time to time. Hence we may be interested in knowing how far the changes of prices of commodities affect the living of the people. This can be done with the help of cost of living Index Numbers.

Cost of living index numbers are designed to measure the average change in the cost of maintaining a given standard of living from year to year. The cost of living index number is computed by comparing the prices paid by the consumers of a particular class of people living in a particular region in two different periods of time for a fixed set of goods and services (for a given standard of living) representing their level of living.

Different classes or groups of people consume different types of commodities. Even the same type of commodities are not consumed in the same proportion by different classes of people. Hence separate index numbers are calculated to measure the

effects of changes in the prices of various commodities on the cost of living of different classes of people. Similarly, separate index numbers are prepared for different types of living for the same category of people because of the change in prices or pattern of consumption from place to place.

It should be clearly understood that the cost of living index number does not measure the actual cost of living. It only tells us whether a particular class of people in a particular region have to pay more or less for *a particular standard of living in* a particular time when compared to a particular base period. In other words, it gives the relative changes taken place in the cost of living when compared to the base period. As it is a relative measure compared to a base period of the same place and not a common fixed place, it is not advisable to compare the cost of living index number of two different locations and arrive at a conclusion about the actual cost of living. We can compare the relative changes but not the actuals, since the actuals in the base period in both the places may not be the same.

Construction of cost of living index number

There are five main procedures in the construction of cost of living index number.

1. Deciding the class of people for whom the index number has to be constructed.
2. Choice of base.
3. Selection of commodities.
4. Determination of weights.
5. Collection of retail price quotations.

1. Deciding the class of people for whom the index number is required

First we have to decide for which class of people, for example industrial workers, or agricultural workers or Government servants etc. the index numbers have to be considered. It is very essential to decide this in clear terms. Besides the people, the area should also be clearly decided.

2. Choice of Base

As the index number is a relative measure, we should decide about the base year with which the present period has to be compared. Generally, the base period should be normal period, without any serious effects on the prices due to any abnormal conditions such as scarcity or abundance.

3. Selection of commodities that are to be entered into the Cost of living index numbers

After selecting the base period, we have to consider the various commodities that are to be taken into consideration in our construction of index numbers. Generally, this is being determined by conducting a Family Budget Survey of the concerned class of people. Since a complete survey is not feasible, a sample survey is always adopted for this purpose. The samples are also selected on random basis to avoid bias. The purpose of this survey or enquiry is to find out how much an average family of the particular class spends on different items of consumption. Generally, the expenditure on various items are broadly classified into the following major groups.

i. Food articles; ii. Cloths; iii. Fuel and Lighting;
iv. House rent; v. Miscellaneous.

These major groups are divided into minor groups, and the minor groups will be further divided into smaller groups in such a way that each small group consists a list of commodities coming under that group.

The family budget enquiry will give the following information.

i. The nature, quality and quantity of each of the commodity consumed by the people.

ii. Their retail prices.

iii. The proportion of the expenditure on a particular item to the total expenditure under all the items in the groups.

iv. The proportion of the expenditure of a particular group to the total expenditure of all the groups.

With the help of the above details the commodities to be included in the construction of the cost of living index are listed. The commodities selected should be those generally purchased by the class of people for whom the Cost of Living Index is constructed. After compiling the family budgets, an average budget is drawn up and this will be considered to be the standard for that particular class of people.

Collection of Retail price Quotations

In the construction of cost of living index numbers, we consider only the retail prices and not the wholesale prices of the commodities, since the consumers used to purchase commodities in small quantities in retail. But the collection of retail prices is tedious and difficult. The prices are subject to greater variation from shop to shop in the same locality.

Therefore, special care has to be taken in the collection of prices. We should try to collect the price from the shop from which more people used to buy. Further, the prices should be more representative i.e. the prices for that sort of commodities that are mostly purchased by the people, if there are more sorts of the same commodity. Prices can be collected through specially trained agents after actually observing the transactions. The retail price quotations collected are averaged afterwards to give an average price for each of the items included in the construction of index numbers.

Determination of weights

The relative importance of various items for different classes of people is not the same. Hence the cost of living index number is always weighted with reference to the importance of the commodities. The relative importance of the commodities is decided on the basis of the expenditure incurred on the commodities or the quantity of consumption as reflected in the average family budget.

Construction of Index Number

Generally, Laspeyre's formula is used in the construction of cost of living index number:

$$\text{Cost of living Index} = \frac{\Sigma P_n q_o}{\Sigma p_o q_o} \times 100$$

In this, the weights of the commodities in the base year is taken as the weights.

The price of each commodity in the current year is multiplied by the base year quantity of that commodity and then the value of the commodity is decided. In this manner, the value of all the commodities is worked out and the grand total of the values of all the commodities at the current year price level is then worked out $\Sigma p_n q_o$. In the same way the total expenditure in the base year for all the commodities at the base year price level is also worked out $\Sigma p_o q_o$. This is indirectly available from the family budget survey also. Afterwards, the current year value is divided by the base year value and the ratio obtained will be multiplied by 100 which will be the cost of living index number for the current year. This method is known as Aggregate Expenditure Method.

Instead of aggregate expenditure method we can also adopt the price relatives and the formula used will be as follows:

$$\text{Cost of Index} = \frac{\sum \frac{P_n}{p_o} \times p_o q_o}{\sum \frac{P_o}{P_o} \times p_o q_o} \times 100 = \frac{\Sigma P_n q_o}{\Sigma p_o q_o} \times 100$$

In this process the index number obtained will be a weighted average of price relatives taking the base year expenditure of each commodity as the weight.

In this method, the expenditure under each item is not calculated as in the first method. Instead, the current year price of each commodity will be divided by the base year price of that

commodity and the current year price will be expressed as a ratio or relative. This price relative will then be multiplied by its weight which is equal to the value of that commodity in the base year which is always a fixed one. The sum of the weighted price relatives of all the items will be divided by the total expemditure in the base year $\Sigma p_o q_o$ which is also fixed. This ratio will then be multiplied by 100 to express the cost of living Index as a percentage of the base year index number. In both the cases the result obtained will be the same.

A hypothetical example of construction of Index is given for purpose of illustration.

Double Weighting

Generally, the cost of living index number is computed by a system of double weighting. We have already stated that the commodities considered in the construction of index numbers are broadly divided into five major groups namely i. Food; ii. Fuel and lighting; iii. Clothing; iv. House rent and v. Miscellaneous. First the index number for each group is separately computed. Afterwards the index number of each group is multiplied by the weight of the respective group. The weight of each group is the percentage of expenditure under the group to the total expenditure of all the groups. The total of these weighted indices of all the groups will be the index number.

COMPUTATION OF INDEX NUMBER BASED OR THE WEIGHTED AVERAGE

S.No. of the items	Price in the Base period P_o Rs.	Price in the current period P_n Rs.	Price relative $\frac{P_n}{P_o}$	Expenditure during the base period $P_o q_o$ Rs.	Percentage of the expenditure to total expenditure w	Price relative × percentage of expenditure $\frac{P_n}{P_o} \times w$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	10	15	1.50	300	20	30.00
2	8	10	1.25	240	16	20.00
3	6	9	1.50	120	8	12.00
4	4	5	1.25	90	6	7.50
5	7	14	2.00	60	4	8.00
6	5	6	1.20	90	6	7.20

7	10	12	1.20	90	6	7.20
8	6	9	1.50	75	5	7.50
9	12	15	1.25	45	3	3.75
10	7	21	3.00	60	4	12.00
11	10	15	1.50	120	8	12.00
12	5	10	2.00	60	4	8.00
13	5	6	1.20	60	4	4.80
14	3	6	2.00	30	2	4.00
15	2	3	1.50	60	4	6.00
	Total		23.85	1500	100	149.95

Exercise

(1) Calculate the index numbers.

Base Year Rs.	Current Year Rs.	Base Year Rs.	Current Year Rs.
(i) 20	30	(ii) 10	12
26	39	15	20
30	45	20	25
22	35	25	30
25	36	30	35

(2) Calculate the price index by using Pasache's and Laspeyris form.

Base Year		Current Year	
(i) Price	Quantity	(ii) Price	Quantity
5	5	6	4
6	6	9	5
9	2	15	2
12	1	16	2
8	3	15	5

(3) Write an essay on the construction of Cost of living index numbers.

(4) Define index number. Write an essay on the construction of Price Index Numbers.

(5) What are the different formulae used in the construction of Index Numbers. Discuss the merits and demerits of the index numbers.

(6) What are the different tests that are generally adopted. Discuss their suitability in the case of different formulae adopted.

TAMILNADU TEXTBOOK SOCIETY

LEARN THOROUGHLY